The Bell System Technical Journal

Vol. XIV *April, 1935* *No. 2*

Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires

By R. N. HUNTER and R. P. BOOTH

When wires are close to each other as they are in cable, the mutual inductance coupling between pairs is not a simple number, constant at all frequencies. Because of the non-uniform and non-symmetrical distribution of current over the cross-sections of the conductors the "effective mutual inductance" is of the form $M = M_a + jM_b$ where both M_a and M_b vary with frequency. This paper discusses the results of certain measurements which have been made of the effective mutual inductance between straight wires and between cable pairs over a wide range of frequencies extending up to a million cycles. This is of interest in connection with crosstalk problems in cable carrier telephone systems.

FOR many years it has been recognized that non-uniform distribution of current over the cross-section of a conductor reduces the efficiency of transmission in either power or communication circuits. With direct current or with alternating current of very low frequency the current is distributed almost uniformly. As the frequency increases, the current distribution becomes more and more non-uniform.

If the two conductors of a circuit are remote from each other the high-frequency current distribution in either conductor is practically symmetrical with respect to its center, the density of the current being lowest in the center of the conductor and highest near the surface of the conductor. If, however, the conductors are close together, the high-frequency current distribution in either conductor is unsymmetrical due to the proximity of the other conductor. This is known as the proximity effect.

It is probably not well known that this proximity effect may have an important bearing on the crosstalk between communication circuits.[1] While the effect is negligible in open-wire circuits, it is quite marked in cable circuits. This paper describes an investigation of the influence

[1] This effect was mentioned in the Carson-Hoyt paper on "Propagation of Periodic Currents Over a System of Parallel Wires," in the *Bell System Technical Journal* of July, 1927.

of the proximity effect on crosstalk between long non-loaded cable circuits which are being studied in connection with the development of high frequency carrier systems suitable for toll telephone cables. More specifically, the paper covers tests made to determine the influence of the proximity effect on the mutual inductance between circuits; data are given both for the case of two isolated non-twisted pairs and for the case of pairs in a quadded 19-gauge cable.

In cable carrier systems it is not practicable to operate like frequency bands in opposite directions on different pairs in the same cable without heavy shields between the pairs. The relatively large level differences that may exist between pairs transmitting in opposite directions would result in excessive crosstalk of the near-end type. Like carrier frequency bands are, therefore, transmitted in the same direction in a cable and the crosstalk between pairs used for carrier systems is of the far-end type.

It has been shown [2] that far-end crosstalk at carrier frequencies between long non-loaded cable pairs can be considerably reduced by the use of simple networks connected between the two pairs at one point in their length. The crosstalk balanced out by such networks is of the "transverse" [3] type. Crosstalk of the interaction type varies in a complicated way with frequency, and cannot, therefore, be annulled by a simple network. For any two similar circuits all the elements of transverse crosstalk, due to the unbalances occurring at various points along the line, arrive at the same time at the far end of the line. The crosstalk currents due to unbalances of the same type such as capacitance unbalances arrive in the same or opposite phase (if the circuits are perfectly smooth). It will be seen, therefore, that a properly designed network connected at one point in the line may be used to practically annul the far-end transverse crosstalk. In order to design the most effective type of network for balancing transverse crosstalk it is necessary to know the manner in which the crosstalk coupling in any elementary length varies with frequency.

The crosstalk coupling between two pairs in an elementary length may be represented by a mutual admittance and a mutual impedance. It can be considered that the voltage between the two wires of the disturbing circuit drives crosstalk currents into the disturbed circuit through the mutual admittance. The currents in the disturbing circuit acting through the mutual impedance also cause crosstalk

[2] As discussed in the Clark-Kendall paper on "Carrier in Cable" in the *Bell System Technical Journal* of July, 1933.

[3] The various types of crosstalk are discussed in the paper on "Open-Wire Crosstalk" by A. G. Chapman in the *Bell System Technical Journal* of January and April, 1934.

currents. The mutual admittance is due almost entirely to capacitive coupling, the leakance ordinarily being negligible in its effect on crosstalk coupling. This capacitive coupling varies but little with frequency and its effect on crosstalk may be balanced out by means of a simple condenser. If the proximity effect were negligible, the mutual impedance would be substantially that of a simple mutual inductance constant with frequency. The crosstalk due to this coupling would, therefore, be balanceable by means of a simple inductance coil. If, however, the proximity effect is not negligible the mutual impedance is due to a complex mutual inductance both of whose components vary considerably with frequency. This is the case in cable circuits and a complex balancing unit must be designed if the complex magnetic coupling is to be accurately simulated.

The mutual impedance, Z_M, between two circuits is by definition the negative ratio of the induced series voltage [4] (e) in the disturbed circuit to the current (I) in the disturbing circuit. Thus,

$$Z_M = -\frac{e}{I}.$$

Since the induced voltage is proportional to the time variation of the magnetic field set up by the disturbing current, it is important to visualize how this field may be altered by changes in the distribution of the current, I, over the cross section of the disturbing conductor. Four types of current distribution will be considered and the effects on Z_M noted.

In order to simplify the following qualitative explanation of the effect of current distribution on mutual impedance it will be assumed that in all cases the disturbed wire is a filament. When the disturbed wire is finite in cross section the effect is generally similar, but more complicated.

Case I—Current Concentrated in a Filamentary Disturbing Wire

In the case of a wire of infinitely small cross section the magnetic field due to a sinusoidal current, I, induces a voltage in another filamentary wire located in this field as expressed by the familiar equation

$$e = -j\omega MI.$$

The mutual impedance is a pure reactance equal to $j\omega M$, where M, the coefficient of mutual inductance, is a pure number and independent of frequency.

[4] This voltage is defined as being the negative of the value of an inserted electromotive force such as to bring the total current in the disturbed circuit to zero.

Case II—Current Uniformly Distributed in a Solid Cylindrical Disturbing Wire

Consider next the case where the total disturbing current is uniformly distributed over the cross section of a solid cylindrical wire. Such a distribution exists exactly with direct current only, but is closely approximated at very low frequencies. Since the magnetic field outside of a conductor carrying a uniformly distributed current is the same as would exist if the total current were concentrated in the center filament, the total induced voltage in a filamentary wire located in this field is again equal to $-j\omega MI$, where M is the same as in the case of two filaments similarly located in space.

Case III—Current Symmetrically Distributed in a Solid Cylindrical Disturbing Wire

The a.-c. distribution in a solid cylindrical wire is not uniform. However, when the wire is at a considerable distance from its return, the current distribution is practically symmetrical about the axis of the wire although its density varies from a minimum value at the center to a maximum value at the surface. Such a distribution is caused by the fact that the counter-electromotive force induced in a filament near the center of the wire due to the current in all of the other filaments is greater than that induced in a filament at the surface. This is the well-known skin effect.

In this case the total current, I, may be considered as distributed in infinitely thin concentric rings in any one of which the current is the same in phase and magnitude at all points. Since the field outside of one such ring is the same as would exist if all of the ring current were concentrated in a filament at the center, the total field due to the sum of the currents in all the concentric rings is the same as would exist if the total current were concentrated at the center of the wire. Thus, the total voltage induced in a filamentary wire by the field set up by a symmetrically distributed current in the disturbing wire is again expressed by $-j\omega MI$, where M is again a pure number as in the case of two filaments.

Case IV—Current Unsymmetrically Distributed in a Solid Cylindrical Disturbing Wire

If a solid wire and its return are placed close together, as in a cable pair, the a.-c. distribution is neither uniform nor symmetrical about the axis of the wire. In this case the magnetic field set up by the current in the return wire contributes to the counter-electromotive

force acting in each filament of the other conductor and causes a further redistribution of the current in that conductor over and above that due to the above mentioned skin effect. This additional alteration in current distribution is known as the proximity effect.

The resultant current distribution can no longer be symmetrical about the axis of either wire. The current in the return wire sets up greater back-electromotive forces in the filaments of the other wire which are close to it than in the more remote filaments. These back-electromotive forces tend to act in opposition to those set up by the current in the wire itself since the current in the return wire is opposite in sign. Hence, the proximity of the return wire reduces the counter-electromotive force acting in the filaments closest to it in the other wire more than it does in the filaments farther away. This results in higher current density in the sides of the wires adjacent to each other.

The current distribution due to the combined action of skin and proximity effects is shown for a pair of round copper wires in space in Figs. 1-A and 1-B.[5] The wires are No. 19 A.W.G. and are separated a distance equivalent to that between wires in 19-gauge cable pairs. The current distribution at 56 kilocycles is shown in Fig. 1-A and at 112 kilocycles in Fig. 1-B. It is seen that the tendency at the higher frequencies is for the current to concentrate on the sides of the wires adjacent to each other. With perfect conductors the current would all be on the surface of the wires and for this wire spacing would be distributed as shown in Fig. 1-C.[6] With actual conductors this distribution is approached as the frequency increases toward the highest conceivable wire communication frequency.

In addition to this unsymmetrical distribution of current with respect to magnitude the currents in various filaments in the conductor may be considerably out of phase with the current at the center. This phase shift may be quite unsymmetrical as indicated for three wire diameters in Figs. 2-A and 2-B. While similar phase shifts occur when only skin effect is present, such shifts are symmetrical about the center of the wire so that the currents at all points in a thin concentric ring have the same phase. Figure 2-A shows the phase shift at 56 kilocycles and Fig. 2-B the phase shift at 112 kilocycles. It is seen that the tendency at the higher frequencies is for the currents at different points on the surface to become in phase with each other. At infinite frequency the surface currents would be in phase.

[5] The current distribution and phase change at 56 and 112 kilocycles were computed from formulas given by Harvey L. Curtis in Bureau of Standards Scientific Paper No. 374, entitled, "An Integration Method of Deriving the Alternating Current Resistance and Inductance of Conductors."

[6] This distribution was calculated by Ray S. Hoyt.

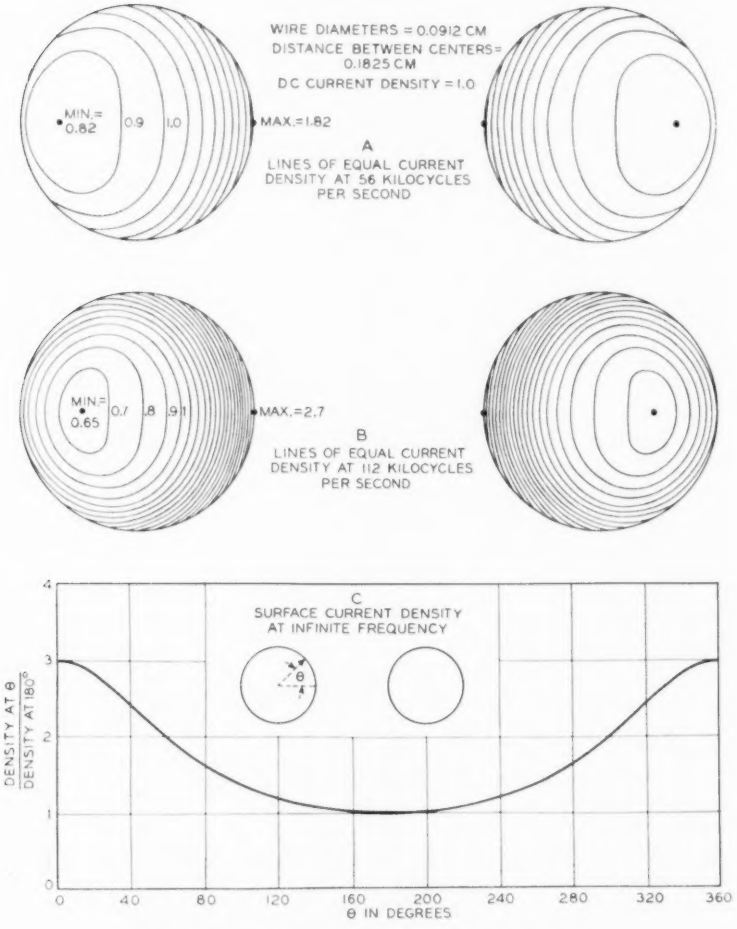


Fig. 1—Current distribution in parallel 19-ga. wires.

If a wire is carrying a total current, I, distributed unsymmetrically in phase and magnitude as in Figs. 1 and 2, the magnetic field surrounding the wire can no longer be the same as would be produced by the same current flowing in the center filament of the wire. The voltage induced in a disturbed filamentary wire located in this field must therefore be different from that induced by the field set up by any of the preceding types of current distribution in the disturbing wire. In each of those cases the induced voltage, e, was exactly in

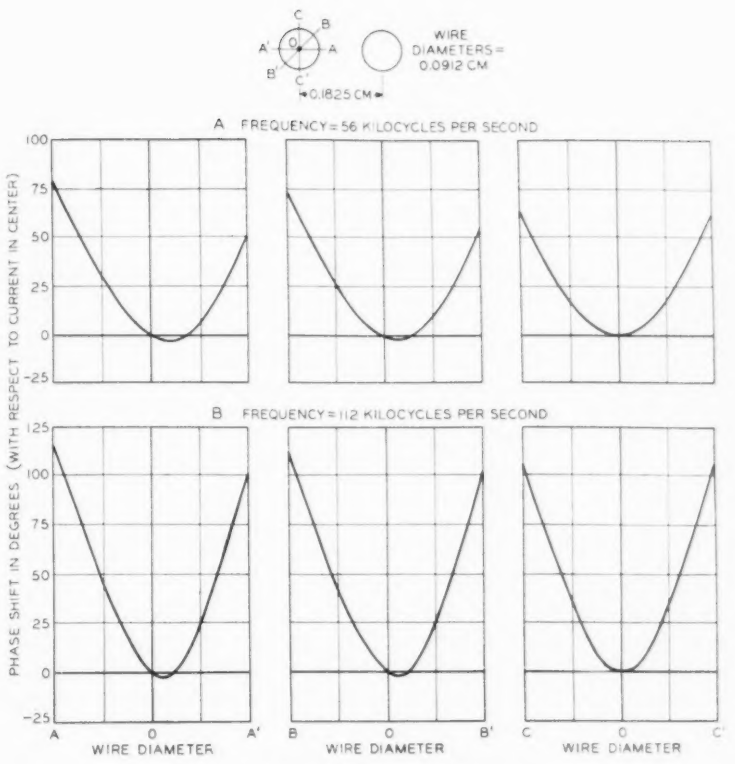


Fig. 2—Phase shift in parallel 19-ga. wires.

phase quadrature with the total disturbing current, I, and the mutual impedance was equal to $j\omega M$, a pure imaginary. For unsymmetrical current distribution in the disturbing conductor, the following discussion shows that the induced voltage can no longer be exactly in phase quadrature with the disturbing current and that the mutual impedance is complex.

The total voltage induced in a disturbed filamentary wire by the total current flowing in a solid wire is the vector sum of the induced voltages due to all of the currents in the various filaments of the disturbing wire. Thus, if $i_1, i_2, i_3, \cdots i_n$ are the vector currents in the various filaments of the disturbing wire and if $m_1, m_2, m_3, \cdots m_n$ are the corresponding coefficients of mutual inductance between each of these filaments and the disturbed filamentary wire, the total induced

voltage in the disturbed circuit is

$$e = -j\omega(m_1 i_1 + m_2 i_2 + m_3 i_2 + \cdots m_n i_n).$$

The mutual impedance, Z_M, may therefore be written

$$Z_M = -\frac{e}{I} = \frac{j\omega(m_1 i_1 + m_2 i_2 + m_3 i_3 + \cdots m_n i_n)}{I},$$

where $I = i_1 + i_2 + i_3 + \cdots i_n$. This is a general expression and holds for any type of current distribution in the disturbing conductor.

In the case of symmetrical current distribution (Case III) all filamentary currents having the value i_1 lie in a ring concentric with the center of the wire. The voltage induced in a disturbed filamentary wire due to all the currents in one such ring is the same as if their total value, I_1, was concentrated in the center of the ring. This voltage is equal to $-j\omega M_1 I_1$ where M_1 is the coefficient of mutual inductance between the center filament of the disturbing wire and the disturbed filamentary wire. The same reasoning holds for currents having values i_2, i_3, $\cdots$ i_n and the mutual impedance may be written

$$Z_M = -\frac{e}{I} = \frac{j\omega(M_1 I_1 + M_2 I_2 + M_3 I_3 + \cdots M_n I_n)}{I}.$$

But $M_1 = M_2 = M_3 = \cdots M_n = M$ since all are computed from the center filament in the disturbing wire to the disturbed filamentary wire. Then

$$Z_M = j\omega M \frac{(I_1 + I_2 + I_3 + \cdots I_n)}{I}$$
$$= j\omega M$$

since $I_1 + I_2 + I_3 + \cdots I_n = I$. This is the same expression for Z_M as given in the discussion on symmetrical current distribution.

However, when the current distribution in the disturbing wire is unsymmetrical in phase and magnitude it is impossible to make the above simplifications. In the general expression

$$Z_M = -\frac{e}{I} = \frac{j\omega(m_1 i_1 + m_2 i_2 + m_3 i_3 + \cdots m_n i_n)}{I}$$

there is no correspondingly simple way to separate the m's from the i's in the complex expression in brackets and the phase angle of the expression may be quite different from that of I. Therefore, e cannot be in phase quadrature with respect to I. In order to put the equation

for Z_M in the same form as in preceding cases the bracketed expression may be arbitrarily rewritten as

$$m_1 i_1 + m_2 i_2 + m_3 i_3 + \cdots m_n i_n = I(M_a + jM_b).$$

Then,

$$Z_M = j\omega(M_a + jM_b) = -\omega M_b + j\omega M_a,$$

where the mutual inductance is now considered complex and as having two components such that

$$M = M_a + jM_b.$$

The total current in the disturbing circuit acting through the component M_b of the mutual inductance sets up an induced voltage in the disturbed circuit in quadrature with the induced voltage due to M_a, and in the same or opposite phase as the total current in the disturbing circuit. Ordinarily the phase will be opposite and the actual values of M_b will be negative with respect to M_a.

Both M_a and M_b vary with frequency. While the total current in the disturbing wire is assumed constant, the unsymmetrically distributed currents in the various filaments change in relative magnitude and phase as the frequency changes. At very low frequencies the current is distributed nearly uniformly in phase and magnitude over the cross-section of the wire. The mutual inductance between this wire and the disturbed filamentary wire is nearly the same as the d.-c. value since M_a cannot be appreciably changed from the d.-c. value and M_b must be very nearly zero. At very high frequencies the major part of the current flows unsymmetrically on the surface of the disturbing wire but the filamentary surface currents are practically in phase with each other. This results again in a low value of M_b because the total induced voltage will be practically in phase quadrature with the total disturbing current. However, due to the unsymmetrical current distribution, the value of M_a is considerably altered from its d.-c. value. At intermediate frequencies the current distribution lies between these two extremes and produces corresponding values of M_a and M_b. Since M_b is zero for both zero and infinite frequency it is evident that a maximum value must be reached at some intermediate frequency.

As noted at the outset of this discussion, the disturbed circuit is assumed to be a filament. In all practical cases the wires involved are finite in cross section, and the reasoning outlined above must be applied to each filament of the disturbed conductor in order to get the total effect.

Discussion of Test Results

In order to study the variation with frequency of the mutual inductance between cable circuits, measurements were made on various combinations of pairs in a 55-foot length of No. 19 A.W.G. toll cable. To obtain information on the performance of the measuring apparatus, measurements were also made on the calculable case of two non-twisted pairs, six feet in length (approximately). Various separations between the two wires of a pair and three different wire gauges were used to show the change in mutual inductance for various degrees of proximity effect.

The measurements were made with a "crosstalk bridge" or admittance unbalance measuring set which permits the measurement of crosstalk in both phase and magnitude. Although the mutual inductance between two pairs may be determined from either near-end or far-end crosstalk tests, it was found that greater accuracy in M_b could be obtained from far-end tests. The computation of M_b from near-end tests involves two terms of opposite sign and of nearly the same magnitude. Consequently a small error in the reading of the crosstalk bridge may result in a considerably greater error in M_b.

The results of the tests on the six-foot non-twisted pairs are shown in Figs. 3 to 9. The data cover a range of 1 to 1000 kilocycles.

In Figs. 3 and 4 the variation with frequency of M_a and M_b is shown

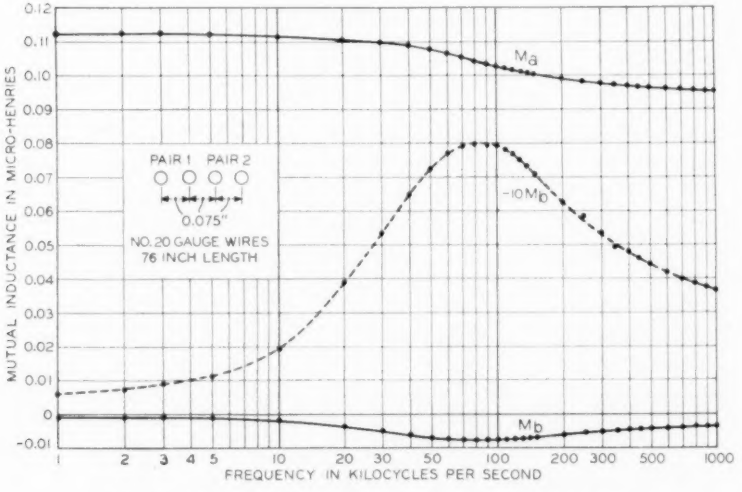


Fig. 3—Mutual inductance between pairs of parallel wires.

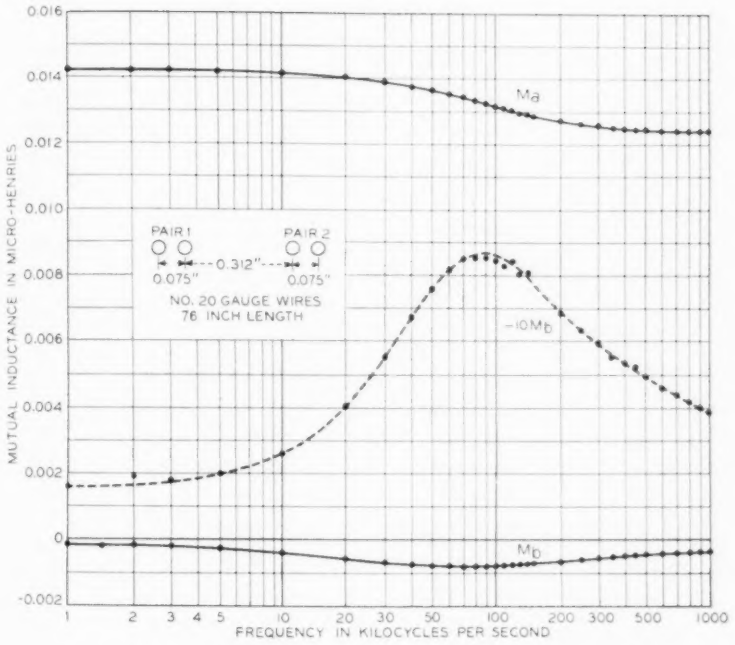


Fig. 4—Mutual inductance between pairs of parallel wires.

for two arrangements of pairs in a horizontal plane. In both cases the axial separation of the two wires of a pair was about 0.075 inch, but in Fig. 3 the axial separation between the nearest wires of the two pairs was 0.075 inch and in Fig. 4 it was 0.312 inch. The wires were No. 20 A.W.G. cotton-covered and were pulled taut to maintain accurate spacing. Two plots are shown for M_b, one actual and the other after multiplying by -10 to show the values more clearly.

The above data are replotted in Fig. 5 to show the frequency variation of M_a and M_b in terms of the values of M_a at one kilocycle. The fact that the frequency characteristics for the two cases are so nearly alike despite the difference in the magnitude of the coupling indicates that the effect depends primarily on the spacing between the wires of a pair and not so much on the relative positions of the pairs.

The proximity effect may be reduced by separating the wires of each pair as shown in Fig. 6. In this case the frequency characteristic of M_a is nearly flat and M_b is so small that it could not be plotted on the same scale as M_a; the curves shown are M_a and $100M_b$. A comparison

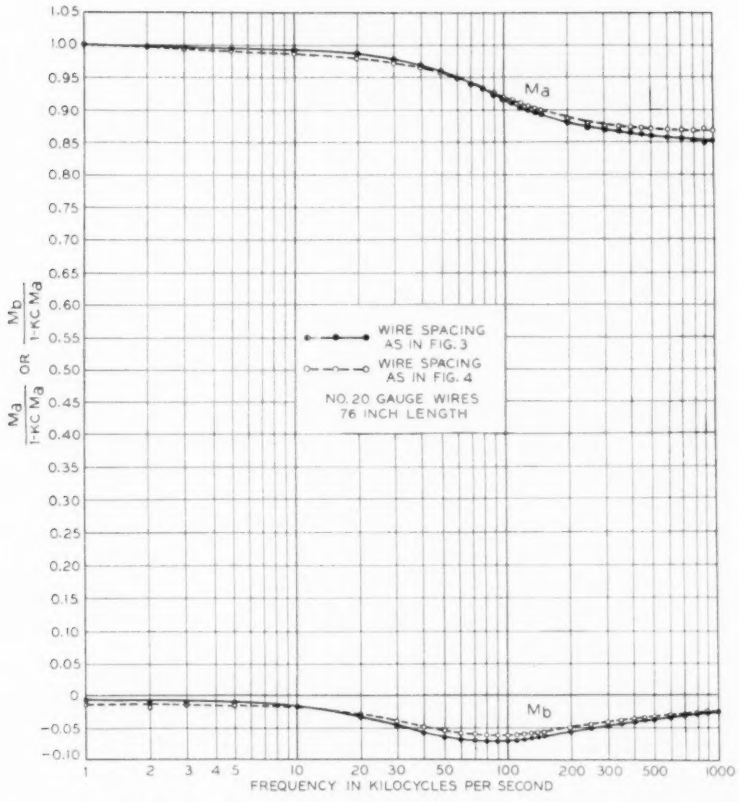


Fig. 5—Mutual inductance between pairs of parallel wires in terms of value for M_a at one kilocycle.

of the curves of M_a and M_b in Figs. 3 and 4 with the curves of Fig. 6 illustrates the relative importance of the proximity effect on magnetic crosstalk in cable pairs and in open-wire pairs. In cable pairs the wires are close together as in Figs. 3 and 4, while in open wire the separation is much greater than that shown in Fig. 6.

The effect of wire gauge on the variation of M_a and M_b is shown in Figs. 7, 8, 9-A and 9-B. Two gauges of wire (No. 10 and No. 18 A.W.G.) were used with centers located at the corners of a 0.14-inch square. The mutual impedance between the vertically adjacent pairs was measured. The corresponding values of M_a and M_b are shown in

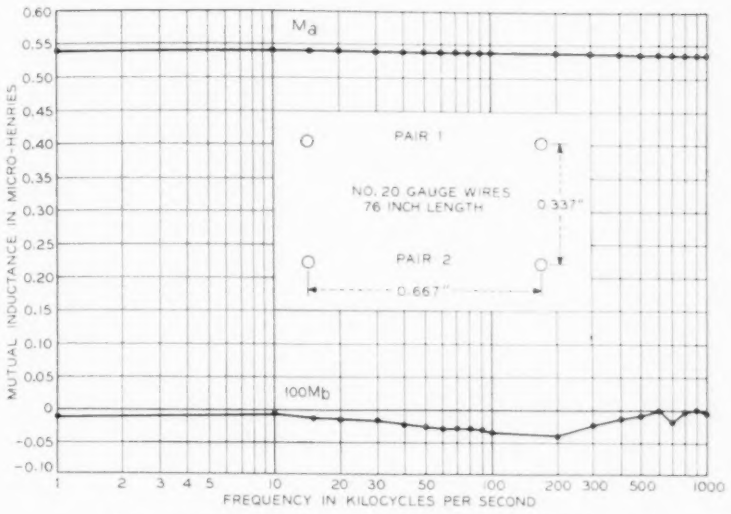


Fig. 6—Mutual inductance between pairs of parallel wires.

Fig. 7 for the 10-gauge wires and in Fig. 8 for the 18-gauge wires.[7] The values of $-10M_b$ are also plotted in Fig. 7 and $-100M_b$ in

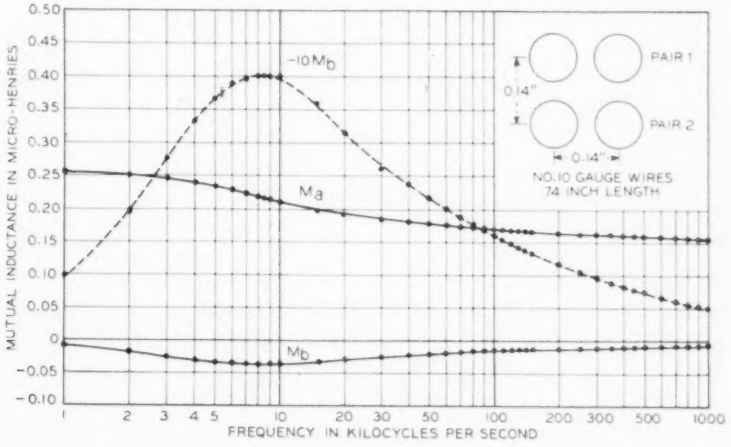


Fig. 7—Mutual inductance between pairs of parallel wires.

[7] These measured values are checked very closely by values calculated by Sallie Pero Mead from the complicated theoretical considerations involved in even the simple case of straight wires.

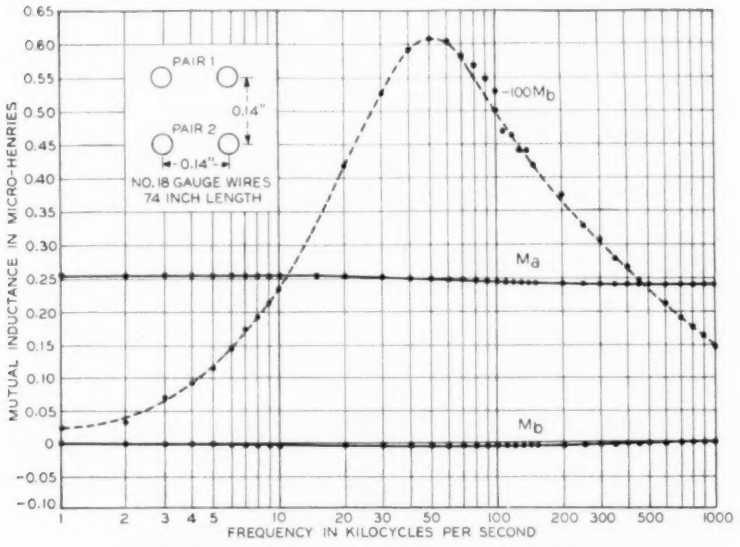


Fig. 8—Mutual inductance between pairs of parallel wires.

Fig. 8 to show the shapes more clearly. A comparison of the frequency characteristics of M_a and M_b for the two wire gauges is shown in Figs. 9-A and 9-B in terms of the one-kilocycle value of M_a for each case. In Fig. 9-A the frequency scale is logarithmic and in Fig. 9-B is linear. As would be expected, the use of smaller wires (No. 18 gauge) decreases the effect of proximity on the magnetic coupling and shifts the frequency at which M_b reaches a maximum value.

The results of tests on the 55-foot length of No. 19 A.W.G. quadded cable are shown on Figs. 10-A and 10-B. The data cover a range of 10 to 480 kilocycles. These figures show the variation with frequency of the average values of M_a and M_b in terms of the ten-kilocycle average value of M_a which is taken as unity. In Fig. 10-A the frequency scale is logarithmic and in Fig. 10-B it is linear. It will be seen that M_b is negative with respect to M_a as in the case of two pairs in space. As in Figs. 3, 4 and 7, curves are given both for M_b and for $-10M_b$, the purpose of the latter curve being to show the shape of M_b more clearly. The value of M_a decreases with frequency, becoming nearly constant above 300 kilocycles at a value 22 per cent less than the value at 10 kilocycles. The component M_b is of negative sign and at 56 kilocycles reaches a maximum value which is 13.4 per cent of M_a at this frequency.

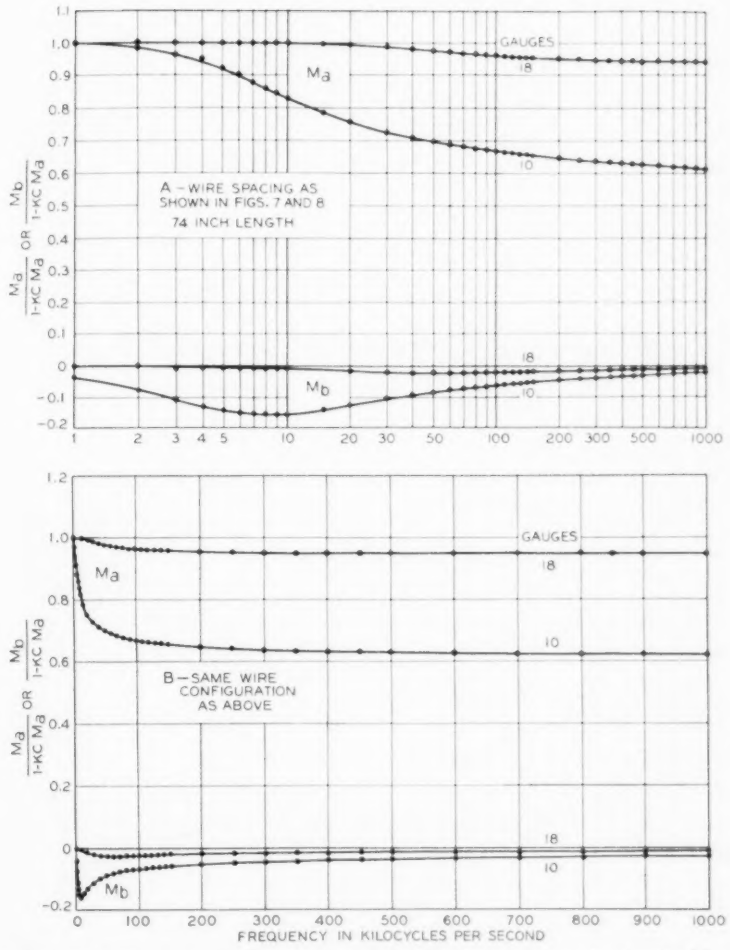


Fig. 9—Mutual inductance between pairs of parallel wires in terms of value for M_a at one kilocycle.

The frequency characteristics of M_a and M_b for individual pair combinations are about the same as shown on Fig. 10, although there are occasional differences such as a positive value of M_b or a change in sign in M_a or M_b at some frequency. Values for pairs in the outside layer did not appear to be much affected by eddy currents in the sheath. As in the case of measurements on parallel wires in space the values of M_a and M_b are very small. For example, between two pairs

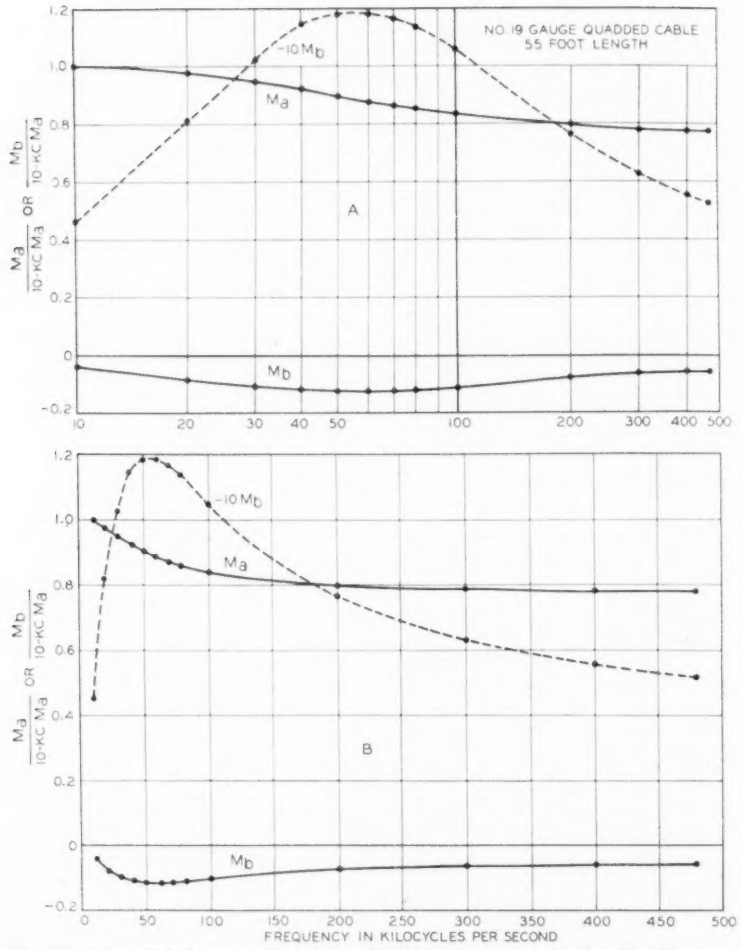


Fig. 10—Mutual inductance between cable pairs in terms of value for M_a at ten kilocycles.

in the same quad the average values at 10 kilocycles are 0.056 and − 0.0030 microhenries. For non-adjacent pairs the values are, of course, much smaller.

Acknowledgment

In this work the writers received much help from Ray S. Hoyt in the matter of general circuit theory. As previously noted, the accuracy of the measurements was established by the work of Sallie Pero Mead in developing a calculation formula for the case of straight wires.

Experiments with Directivity Steering for Fading Reduction *

By E. BRUCE and A. C. BECK

Short-wave fading is largely due to phase interference between multiple path signals of varying path length. Fortunately, stable angular differences usually exist between these paths at the point of reception. It is therefore desirable to employ antenna directivity which is "steerable" and sufficiently sharp to accept only one of the several paths in order to reduce this fading.

This paper describes experiments made with a "steerable" directive antenna during reception of transoceanic short-wave signals. The results demonstrate that sharp angular discrimination is a basically sound method of combating fading which is due to phase interference.

Introduction

RAPID fading in radio communication has been recognized for some time as being due to the interaction of distinct components having different transmission times. The possibility that these components might arrive from slightly different directions was suggested by various observed facts, among which was the behavior of sharply directive antennas.

It has been noticed in the past that fading was affected by the directivity of the receiving antenna.[1] An example is given in the oscillograph records of Fig. 1 showing observations made by the authors some years ago at Cliffwood, New Jersey. These illustrate a condition of less fading on a large "inverted vee" [1] antenna than on a small non-directional antenna, using telegraph signals received from station GBK in England. Beating the signal with a local oscillator provided the audio frequency which was recorded. The directive antenna output was recorded on the upper trace while the lower strip recorded the output of the substantially non-directive, comparison antenna.

Such observations as these suggest the possibility of controlling and reducing fading by a systematic use of sharp directivity. The present paper reports some experiments in which changes in fading are correlated with changes in the directive pattern of a rhombic antenna [1] made by mechanically changing its shape.

It may be reasoned that, where the total differences in the path

* Published in April, 1935 issue of *I. R. E. Proc.* Presented at meeting of I.R.E., April 3, 1935.

[1] E. Bruce, "Developments in Short-Wave Directive Antennas," *Proc. I. R. E.*, Vol. 19, pp. 1406–1433, August, 1931; *Bell Sys. Tech. Jour.*, October, 1931.

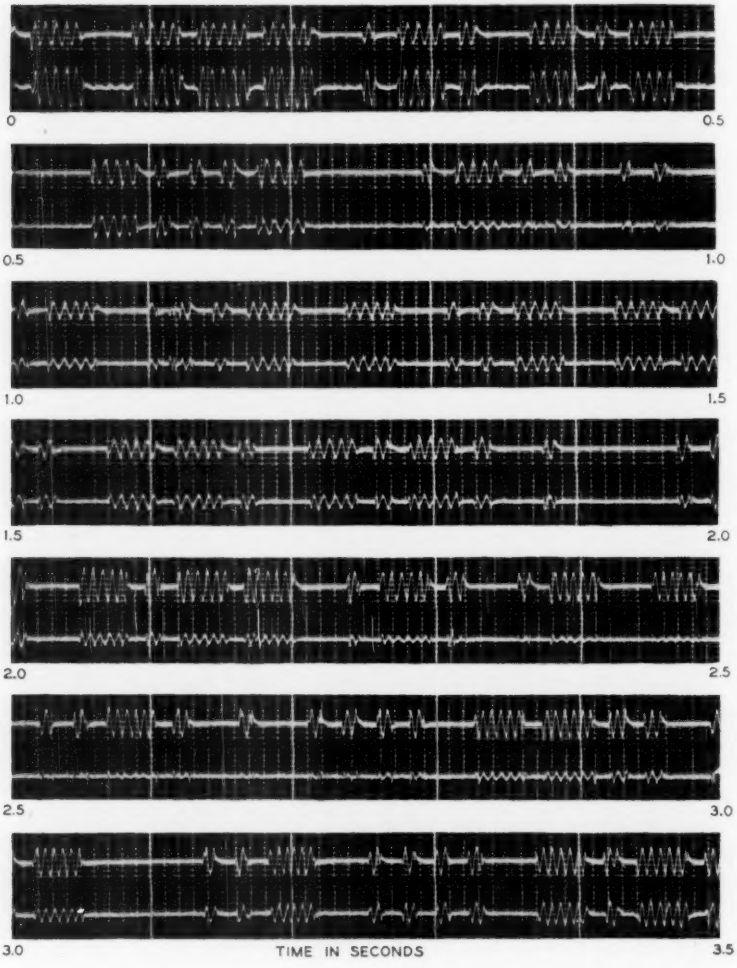


Fig. 1—Oscillographic record of carrier fading reduction. The upper trace is proportional to the output of a large "inverted vee" antenna, and the lower trace to the output of a half-wave vertical antenna when receiving station GBK on 16.6 meters. Taken at Cliffwood, N. J., November 16, 1927, 4:00 P.M., E. S. T.

lengths are small, variations can result only in the carrier and side bands fading in and out together or in other words "general" fading. In such cases, there either may or may not be appreciable angular separation between the multiple waves at the point of reception. However, there is little question that, where multiple waves cause a "selective" fade over a speech band which is, of course, a very small percentage of the carrier frequency, a material path length difference must exist. Where this is the case, it is difficult to conceive of wave routes which do not possess appreciable angular separation between them at the place of reception. The truth of this latter point is of vital importance in this discussion.

The hope of success in fading reduction through directivity rests on the possibility of a continuous, stable angular separation between the interfering waves during times when fading is really troublesome. Fortunately this possibility is reasonably existent; therefore it should be possible to reject all but one of the interfering paths, by means of sharp directivity, with a consequent reduction in selective fading.

Description of Equipment

Tests have shown [2] that a greater degree of angular spread between the multiple waves exists in the incident vertical plane than in the horizontal plane. It might be expected, then, that such a scheme as that illustrated in Fig. 2 would be worth trying. Here the steep edge

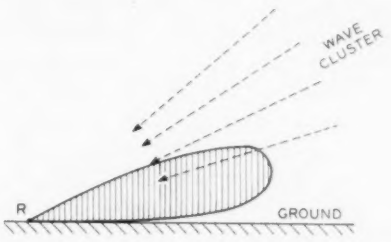


Fig. 2—Edge system for achieving fading reduction with moderate antenna directivity.

of a moderately sharp directional characteristic is moved just far enough into the wave cluster, assumed directively stable, to accept the first wave. Obviously it is possible to approach the wave cluster from the bottom as illustrated or we may approach the cluster from above.

[2] H. T. Friis, C. B. Feldman, and W. M. Sharpless, "The Determination of the Direction of Arrival of Short Radio Waves," *Proc. I. R. E.*, Vol. 22, pp. 47–78, January, 1934.

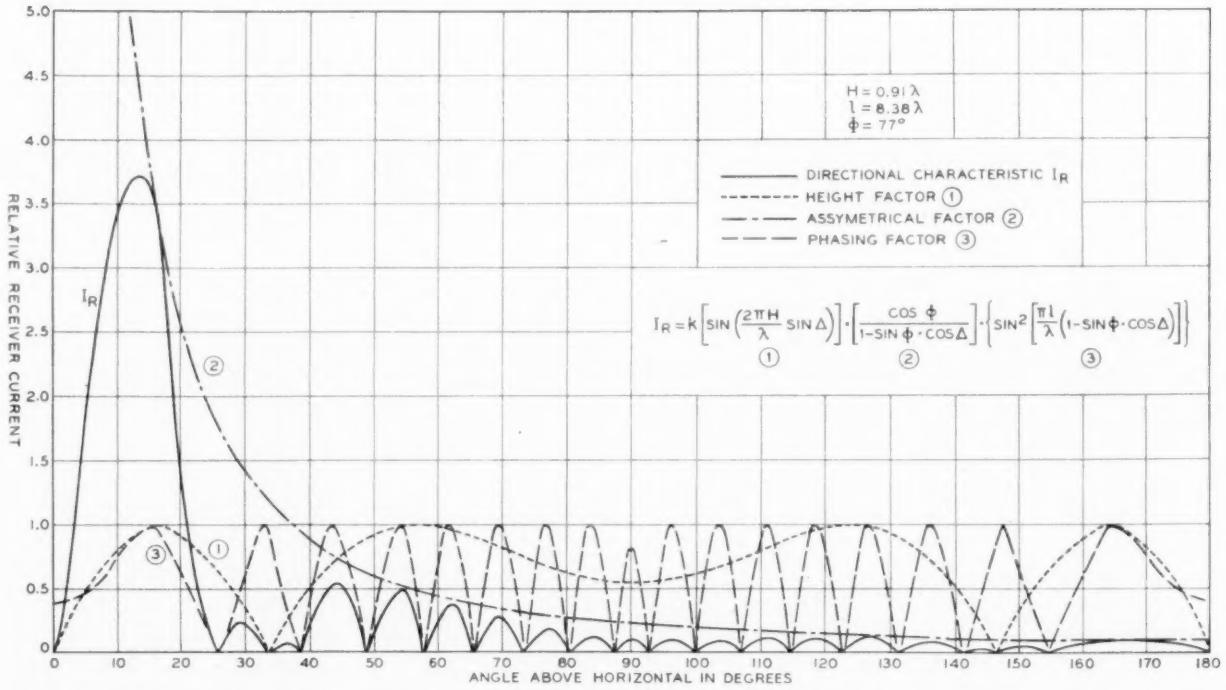


Fig. 3—Receiver current diagram of incident plane directivity for one adjustment of a steerable rhombic antenna. The factors in the equation are plotted separately.

A primary essential in this scheme is that no minor ears of the directive diagram be of appreciable size.

Using this scheme, it is not necessary to discriminate completely against the adjacent waves for practical benefit. A discrimination of ten decibels between two adjacent waves of equal amplitude will make improbable a fade deeper than 5.7 decibels from their sum. Fading of this depth would be relatively unimportant for ordinary speech transmission.

An edge wave may at times be much smaller in amplitude than the adjacent waves. The scheme under discussion may be usefully operative even in this situation since the very smallness of the edge wave means that it cannot be seriously harmful. When signals are weak, the edge of the directive diagram should be advanced until a large amplitude wave is encountered. Some fading of small depth would then exist.

It was stated above that the antenna system used should have no minor ears of appreciable size. At the same time, the edge position of the major loop must be continuously adjustable. A simple method of meeting these requirements is that of mechanically moving the elements of a "long-wire" antenna in space so as to alter the manner of its exposure to the space waves.

Figure 3 is a rectangular plot of the incident plane directive diagram of a large horizontal rhombic antenna when used for GBW on 20.78 meters. The essential antenna dimensions are indicated on that figure as well as the equation for the directive diagram.

Each bracketed quantity in the directive equation of Fig. 3 is separately plotted on that figure together with the final resulting product. Factor 3, known as the "phasing" factor, exerts the greatest influence on the shape of the major lobe. This factor contains only the variables of length l and the angle ϕ, defined as half of the side interior angle. The length cannot be made easily variable but the angle ϕ can be readily adjusted. When an adjustment in ϕ is available for this antenna, Fig. 4 gives the directions of the major lobe maxima, and the first nulls, above the horizontal for a series of wave-lengths. It is evident that a useful degree of steering is provided without limiting the desirable variable wave-length features of the antenna. In all cases, the minor ears remain small.

In Fig. 5 is shown a remote controlled power-winch system for altering the interior angles. This experimental system in slightly modified form was in operation at Holmdel, New Jersey, for some time, without any antenna breakages. This was primarily possible because the angles of flexing were very small and copper-clad steel wire was

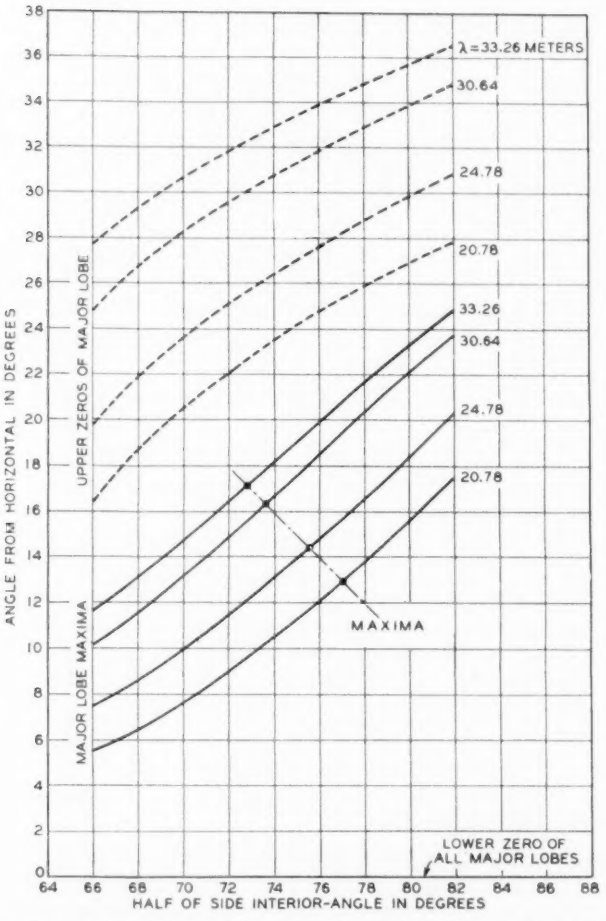


Fig. 4—Steerability, at several wave-lengths, of the horizontal rhombic antenna used for the fading reduction studies. The antenna element lengths were 184 meters and their height 19 meters.

employed in the antenna. The power-winch was equipped with automatic safety stops at the extreme positions, also with a potentiometer which was coupled to the winch to permit the use of a voltmeter as an antenna position indicator. This position indicator was located at the operator's position. By using counterweights, the required size of the winch motor is reduced.

The adopted system for observing selective fading required a fre-

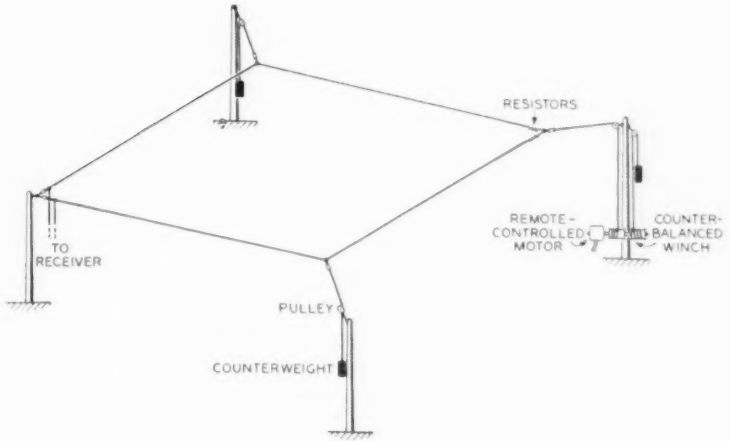


Fig. 5—Mechanical layout of the steerable horizontal rhombic antenna.

quency wobbled carrier from the transmitting station. By beating this frequency-wobbled carrier with a local fixed frequency, a wobbled audio note was obtained after detection. This audio output was impressed on the horizontal plates of a cathode ray tube, after being amplified by an audio amplifier. This produced a horizontal spot deflection on the tube screen which was directly proportional to the field strength of the signal. The vertical plates had a locally adjusted sweep circuit voltage impressed on them to produce vertical spot deflections. The sweep frequency was synchronized with the wobble rate so that the extreme upper and lower deflections occurred at the same instant as the respective upper and lower frequencies of the wobble. Figure 6 indicates the cathode ray picture of a signal without selective fading while that of Fig. 7 shows a severe case of selective fading. It is apparent that general fading was revealed by the horizontal collapse of the rectangle of Fig. 6.

It is an interesting fact that upon the first appearance of the cathode ray figure, with the wobble rates employed, it is a horizontal line moving up and down, but after a few seconds, the traced solid figure stands out clearly, due to the persistence of vision.

One of the surprising results of experience with this system was that, at times of severe fading, eight or ten depressions were occasionally seen within a sweep of a few hundred cycles.

For comparison purposes, there were two complete outfits, as described, with their cathode ray tubes mounted side by side. One outfit

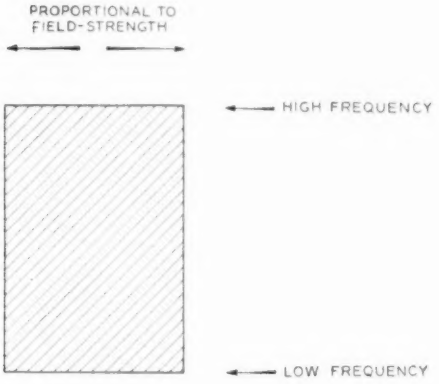


Fig. 6—Cathode ray oscillograph figure for no selective fading when observed with wobbled carrier.

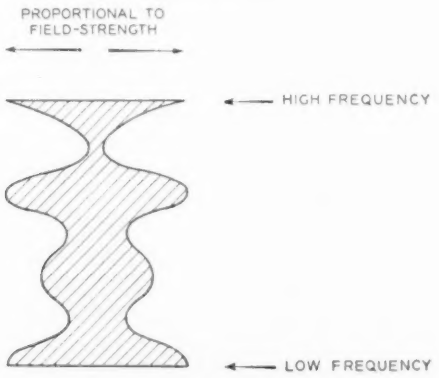


Fig. 7—Cathode ray oscillograph figure for severe selective fading when observed with wobbled carrier.

operated on a simple antenna system, as a standard of comparison, while the second was connected to the adjustable directive antenna. Fig. 10 is a photograph of this apparatus.

Other tests also going on at Holmdel, N. J., were concerned with the measurement of the comparative delay times and the respective angles of the various paths of the waves.[2] To permit this, the British Post Office transmitter sent pulses of very short duration. At the receiving point, a single transmitted pulse frequently appeared as several spaced pulses when a sweep circuit was employed. The spacing enabled the measurement of the relative time delays. It was found to

be the apparently invariable fact that the earlier arriving pulses are the lower in angle with the horizontal and are relatively stable in direction. These tests suggested that a somewhat similar scheme of observations would be useful to the present work since, if pulses were similarly employed, one would actually see the effect on each individual path of steering the antenna.

Accordingly, cathode ray equipment was constructed employing a

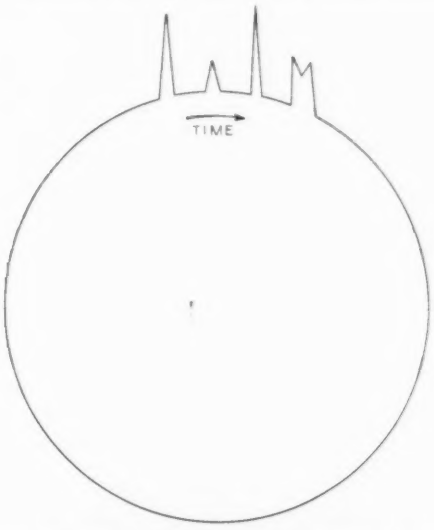


Fig. 8—Cathode ray oscillograph pulse figures when using the circular sweep circuit. The circumference is traversed by the spot in twenty milliseconds.

circular sweep system, in place of the usual linear sweep, thus making the entire time interval always in view. Figure 8 illustrates how the pulses sometimes appeared during this sweep. Since the pulses were always vertical, their definition was lost if permitted to slide down into

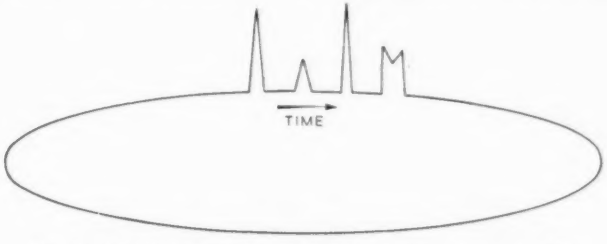


Fig. 9—Cathode ray oscillograph pulse figures when using the elliptical sweep circuit.

the "3 o'clock" or "9 o'clock" positions of the circle. This possibility was considerably reduced by employing the ellipse in Fig. 9 instead of the circle. For general observation purposes the ellipse was used but for more accurate time delay measurements the circle was employed.

The British Post Office station transmitted pulses at intervals of 0.02 second. In order to synchronize with them, an oscillator variable about 50 cycles was used to keep the pulse position stationary. A split-phase circuit feeding the four cathode ray plates produced the circular or elliptical sweep. This equipment is also shown in Fig. 10.

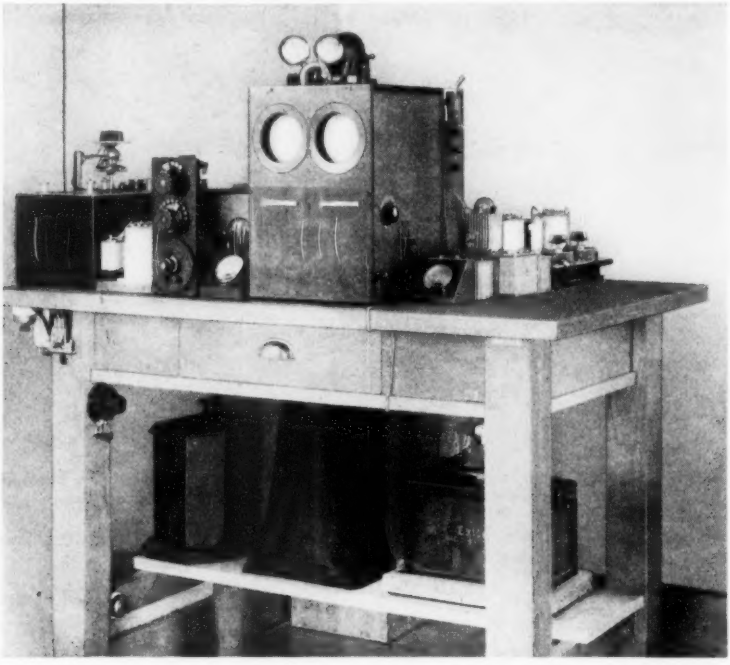


Fig. 10—Cathode ray oscillographs, their amplifiers, and the sweep circuit installation. The meter in the center of the table is the antenna position indicator.

Some studies of general carrier fading were made with a pair of magnetic counters actuated by trigger gas tubes. These fading counters were operated together with automatic recorders so as to maintain the same integrated average signal output. Since, in the recorder integration, ten-second intervals elapsed between gain readjustments, the fading counters operated to record all quick faces, during these inter-

vals, which fell below the average output level by any prescribed amount. A photograph of this equipment is shown in Fig. 11.

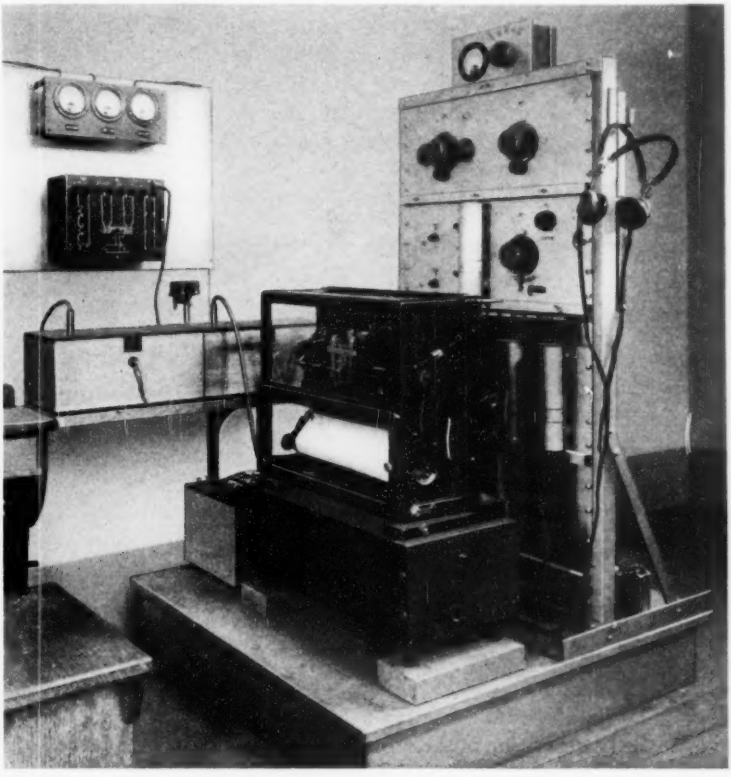


Fig. 11—Field strength recorder and fading counters used for fading reduction studies.

Results

Cathode ray tube observations of selective and also general fading were made on the British Post Office stations GBW and GBU using wobbled carrier. Whenever possible, these observations were made at half-hourly intervals. For record purposes, arbitrary numbers ranging from 0 to 4 were adopted. Zero meant very little fading (five per minute or less) and the most severe cases were represented by 4. These figures were recorded separately for the standard antenna and for the rhombus. The difference between the numbers assigned to each antenna gave an indication of the fading reduction accomplished.

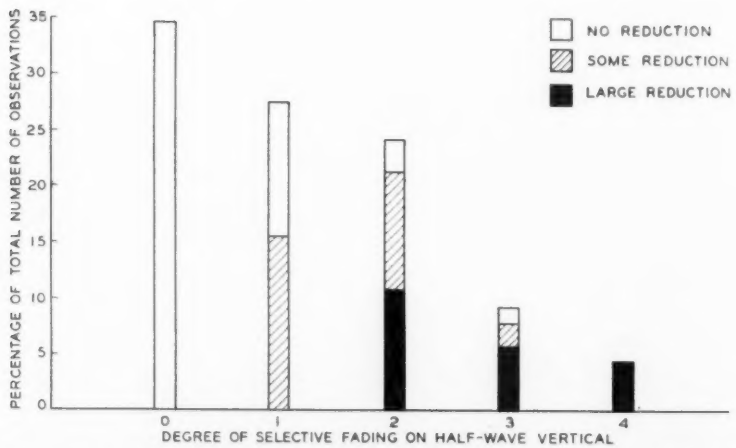


Fig. 12—Selective fading severity and its reduction at the best positions of the rhombic antenna. Stations GBU and GBW, March and April, 1933.

Figure 12 is a summary of results of these half-hourly observations made during the working hours of March and April, 1933. Disregarding the fact that portions of that figure are shaded, the total lengths of the vertical bars represent percentage of the total number of observations plotted against the degree of the selective type of fading, observed on the comparison antenna, as indicated on the abscissas.

During each of the above observation intervals, the rhombus was steered over its available range to determine the best position for reduction in selective fading. Each of the vertical bars in Fig. 12 is subdivided by shading into the various degrees of fading reduction obtainable at the best position of the adjustable rhombus. The solid sections represent large selective fading reductions, the cross-hatched sections are fair reductions, while the unshaded portions indicate that the reductions were not of appreciable magnitude.

Analyzing Fig. 12, the results show that 51 per cent of the readings gave no reduction in selective fading; however, for 35 per cent of the readings there was practically no selective fading to be reduced. On the other hand, if one disregards the rather mild and therefore relatively harmless fading cases, graded 0, 1, and 2, rhombic fading reductions were possible 89 per cent of the remaining time, so that when selective fading on the comparison antenna was really troublesome, it is important to note that an appreciable rhombic selective fading reduction was nearly always accomplished. By deliberately steering the rhombus to a disadvantageous angle, it was possible four per cent

of the time to make the selective fading worse on the rhombic antenna than on the comparison antenna, but no case has been observed where, at an ordinary rhombic antenna setting, the selective fading was not at least equal to or less than that on the comparison antenna.

While the cathode ray tube figures indicated some degree of general fading, where all frequencies fade together, it was evident that this type of fading is of far less importance than the selective type of fading, in fact it was rarely noticeable except when the selective fading was almost absent.

Figure 13 is a photograph of permanent wobble records of selective

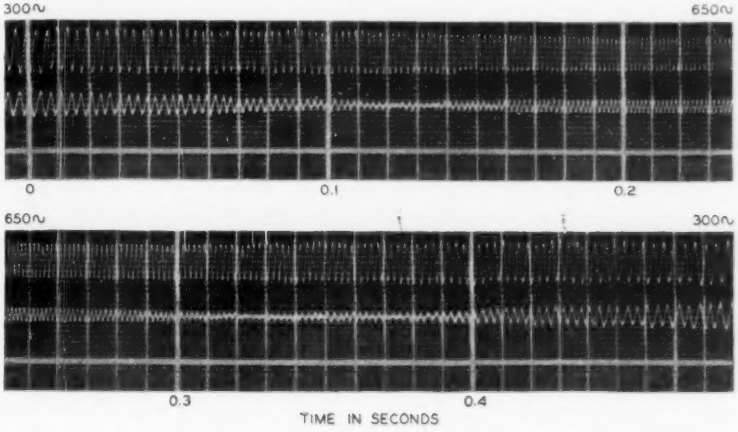


Fig. 13—Oscillographic record of selective fading reduction. The upper trace is proportional to the output of the rhombic antenna, when the angle ϕ equalled 69 degrees, and the center trace is proportional to the output of the half-wave vertical antenna. The lower string was idle. Wobbled carrier from station GBU, April 19, 1933, 4:00 P.M., E. S. T.

fading as recorded by the string oscillograph previously mentioned. The center string was actuated by the signals from the half-wave vertical comparison antenna while the rhombus signal was fed to the upper string. The third string was not utilized. The frequency wobble can be seen on close examination and as each small timing division is 0.01 second, the audio frequency is recorded. The record has been marked at the wobbled frequency extremities.

Figures 14, 15, and 16 are sketches of three interesting series of pulse patterns observed on the rhombic and comparison antennas. The three groups reading from left to right show the effects on the individual pulses of the steering of the rhombus, as indicated by the angle ϕ. The steering achieved at these angles can be seen by referring again

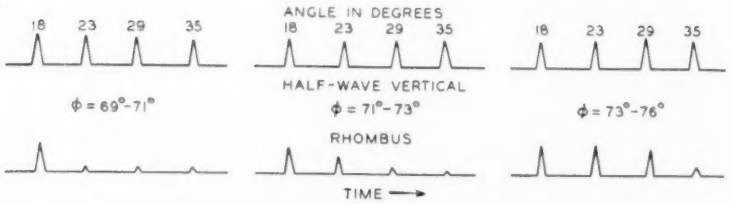


Fig. 14—Pulse pattern changes with steering, March 8 and 9, 1933. Station GCS on 33.26 meters.

to Fig. 4. Marked over the individual pulses are the arrival angles above the horizontal, measured through the cooperation of co-workers.

Figure 14 is of a test, at thirty-three meters, during a period when a wide angular spread of the cluster prevailed. Four narrow pulses of similar magnitude appear on the half-wave antenna. The progressive effect of suppressing the higher angle waves by steering the rhombus is shown. Very appreciable selective fading reductions are possible under such conditions.

Figure 15 is a sketch of twenty-meter observations during a period

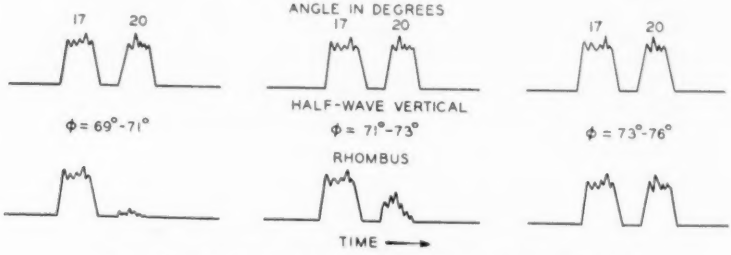


Fig. 15—Pulse pattern changes with steering, April 8, 1933, 2:00 P.M., E. S. T. Station GBW on 20.78 meters.

when selective fading reductions were achieved at the lower angle antenna settings. The broad, flat tops of the pulses are incidentally an interesting contrast to those in Fig. 14. These are possibly due to an increased horizontal spread of wave angles.

Figure 16 is of a case where it was possible deliberately to make the fading on the rhombus worse than that on the comparison antenna. Since the later pulse had a higher amplitude than the earlier one, rhombic steering by equalizing the relative amplitudes, as shown in the left-hand figure, made the selective fading very bad indeed. The opportunities for producing a result of this nature are rather rare, in fact in our previously mentioned wobble studies it was possible to make

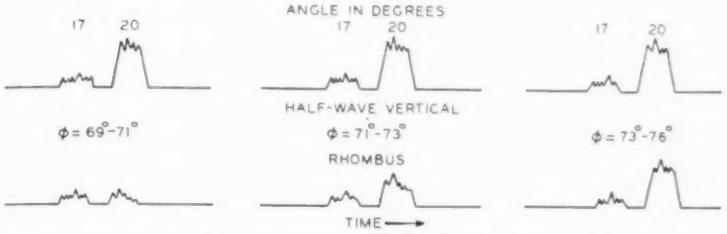


Fig. 16—Pulse pattern changes with steering, April 8, 1933, 2:40 P.M., E. S. T. Station GBW on 20.78 meters.

the selective fading worse only four per cent of the total time of observations.

Occasionally, and in particular on twenty meters, only slight selective fading was observed. When pulse transmissions were available during these times, only one major pulse could be seen. Really bad fading invariably occurs when multiple pulses, which are widely spaced in time, are observed.

It may be evident, from the previous discussions in this paper, that the change in antenna output, with steering, is closely related to the number and spread of the waves arriving and to the selective fading improvements obtainable. Figure 17 shows three cases of results secured by reading relative gain changes, as shown by automatic recorders.

Case 1 is typical of a closely spaced wave cluster arriving at an average angle of about ten degrees above the horizontal. Case 2 can be explained as due to a narrow wave cluster at eleven degrees plus another of less amplitude at eight degrees. We would ordinarily expect annoying selective fading in such an event. Should we deal with many closely spaced waves having a large angular spread, very little gain change would be evident while steering the rhombic antenna, but selective fading improvements over the comparison antenna might still be possible.

Curve 3 is of considerable interest in that it served as one of the experimental checks of the theoretical directive pattern calculations. The change in gain with steering is so well defined that probably only one wave-direction existed. This belief was supported by an absence of noticeable fading. Independent measurements, made by an average angle measuring installation[2] consisting of two horizontal dipoles at different heights which determines the average angle by the ratio of the respective outputs, gave the arrival angle at from eighteen to nineteen degrees above the horizontal. Figure 4 indicates that a ϕ-angle

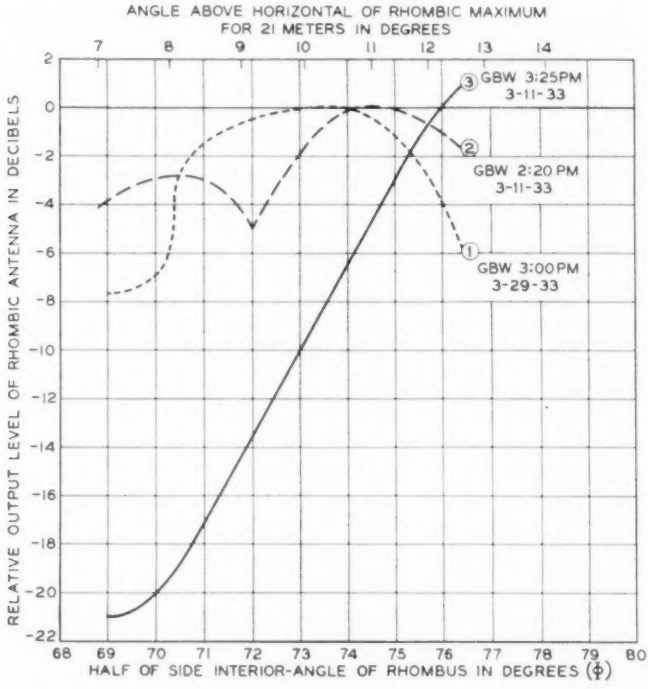


Fig. 17—Horizontal rhombic antenna output changes with steering as shown by automatic field strength recorders. Corrections for changes in signal level with time, as obtained from a half-wave vertical antenna, have been applied.

of about 68 degrees would place a null at this angle. While the range of steering of the rhombic antenna in use did not permit an adjustment to less than about sixty-nine degrees, the trend of the curve leaves little doubt as to the correctness of our null point calculation.

As might have been expected, the previously described fading counters for studying general carrier fading showed that reductions were usually obtained at the directivity positions which also gave the least selective fading. This type of apparatus is incapable of determining whether general fading or selective fading conditions are affecting the amplitude of the fixed carrier frequency.

Conclusion

It is believed that the results, discussed in this paper, demonstrate that sharp angular discrimination is a basically sound method of combating selective fading.

A General Theory of Electric Wave Filters *

By H. W. BODE

THE growth of the field of electric wave filters since their original discovery by Dr. G. A. Campbell shows the filtering action by no means inheres in any particular physical configuration. Filters have been built, for example, as recurrent or non-recurrent ladder structures, as lattices, as bridged-T's, and in a variety of combinations of transformers with ordinary elements. No general theory uniting all these configurations has, however, been developed. Each structure has been treated by methods which are primarily adapted to that configuration alone. In consequence, such questions as the relations between filters of different types and the possibilities of securing improved characteristics by going to a still wider variety of configurations have remained unsettled.

The present paper is an attempt to develop a general filter theory independent of any particular structure, by means of which these questions can be answered. For the sake of a rigorous discussion, the term "filter" has been used to signify a four-terminal network of ordinary lumped elements which, when terminated in its image impedances, transmits freely one continuous band of real frequencies and attenuates all other real frequencies. Since in actual operation the distinction between free transmission and attenuation is always more or less obscured by terminal effects and parasitic dissipation, this definition is necessarily somewhat arbitrary. It agrees, however, with common usage except in its rejection of multiple band-pass filters, which are rarely used in practice.

It follows from the definition that a "filter" can include only reactive elements. Otherwise, however, the structure considered may be an arbitrary four-terminal network and may include transformers as well as ordinary inductances and capacitances. The analysis is based upon a combination of the ordinary image parameter method of analyzing networks and the normal coordinate method familiar in the dynamics of vibrating systems. It is found that the conditions for filtering action can be expressed by means of relations between the set of normal coordinates of the network when it is short circuited at both

* This paper is a summary of a recent article with the same title appearing in the *Journal of Mathematics and Physics* of Massachusetts Institute of Technology, November, 1934. It is included largely for its value in connection with the accompanying paper on "Ideal Wave Filters."

ends and the sets obtained by open circuiting one or both ends. The same normal coordinate solutions also furnish convenient parameters in terms of which general expressions for the image parameters of the structure can be built up. For a band-pass filter, for example, the typical result is

$$\text{Tanh } \theta = k_1 \frac{\sqrt{a_{c_1}} a_{n+1} \cdots a_{p-1} \sqrt{a_{c_2}}}{a_n \cdots a_p},$$

$$Z_I = ik_2 f \frac{a_2 \cdots a_{m-1} \sqrt{a_{c_1}} \sqrt{a_{c_2}} a_{q+1} \cdots a_r}{a_1 \cdots a_m \qquad a_q \cdots a_{r-1}},$$

where a_j symbolizes a frequency factor of the form $1 - \frac{f^2}{f_j^2}$ and

$$0 \leq f_1 \leq f_2 \leq \cdots \leq f_m \leq f_{c_1} \leq f_n \leq f_{n+1} \leq \cdots \leq f_h \leq f_{c_2} \leq f_q \cdots \leq f_r \leqq \infty.$$

The formulæ are almost exactly similar to those familiar in the theory of the lattice, except that the quantities $f_1 \cdots f_n$ which are now natural frequencies of the network as determined under the previousiy described conditions have a different significance. As in the lattice, however, they fall into three groups: $f_1 \cdots f_m$ and $f_q \cdots f_r$, which affect Z_I only; $f_n \cdots f_p$, which affect θ only; and the cut-offs, f_{c_1} and f_{c_2}, which enter into both expressions. The formulæ can be extended to low-pass, high-pass and all-pass structures by allowing the cut-offs to assume the limiting values zero and infinity respectively.

Certain further restrictions upon the image impedance and transfer constant of physically realizable filters may be obtained from the consideration of another system of parameters, $r_1 \cdots r_n$, defined as the roots of the equation $\tanh \theta = 1$. They are usually of either single or double multiplicity. The importance of the roots depends upon the fact that in combination with the cut-offs they are sufficient to determine θ at all frequencies. The restrictions to which they lead may be divided into two sets. The first affects the transfer constant alone and is expressed in terms of limitations on the allowable positions and multiplicities of the roots. The second is concerned with the restrictions which must be placed upon the relation between the transfer constant and the two image impedances. It may be expressed by the statement that when the transfer constant and one image impedance have been chosen as functions of frequency, the second image impedance is determined as a function of frequency to within a constant multiplier. The differences between the two image impedances depend only upon the roots of single multiplicity, so that if only double roots are involved the structure is necessarily symmetrical.

The second half of the paper is devoted to an interpretation of known filter theory in terms of these results and to an attempt to extend this theory until it affords a definite technique for the construction of any filter which the preceding analysis has shown to be physically admissible. The method followed depends upon the fact that when a number of filter structures with matched image impedances are connected in tandem, the roots, $r_1 \cdots r_n$, of the resulting filter will be the aggregate of the roots of all the individual units of which it is composed. This allows us to represent the general filter as a composite structure in which each constituent represents one or at most a few of the total number of roots. The resulting networks are very similar to the familiar Zobel type composite filter, especially when it is noticed that the various required roots can be obtained from simple prototype structures by transformations analogous to the m-derivation, and that the preceding classification into roots of single and double multiplicity corresponds in the composite filter to a classification of the constituent structures into half and full-sections.

In spite of these relations, the usual composite filter theory must be extended in several ways if the general filter is to be adequately represented. The first extension is demanded by the fact that in the general filter we must be able to assign one image impedance characteristic of each of the constituent sections in any form compatible with the preceding general equation. The required image impedances are not obtainable from ordinary ladder structures. When the constituent involved is a double root, or full-section, structure however, the required impedance can readily be realized by resorting to the lattice form. With half-section structures the procedure is more complicated. It is necessary to make use of a combination of Dr. Zobel's multiple m-derivation and a new transformation, described as an h-derivation, which alters the impedances of ladder type half-sections without affecting their transfer constants.

Similar extensions are also needed to provide the requisite variety of transfer constants. Roots falling within certain ranges can be provided by ordinary ladder structures, m-derived, in the usual way, with a real value of m less than one. In order to complete the list, however, it is also necessary to consider single sections derived with real values of m greater than one, which may be realized as lattices or as ladder structures with mutual inductance, and pairs of sections derived with conjugate complex values of m. By including all of these types of sections we can construct physical networks giving any filter characteristics falling within the general limitations discovered in the first part of the paper.

Aside from its purely theoretical interest, the analysis leads to two results of immediate practical value. The first is the introduction of new characteristics by the complex m and h-derived sections. The h-derived structures can be dismissed briefly. They are chiefly of interest for their impedance characteristics, which resemble those found in symmetrical lattices. They allow us, however, to extend these impedances to unsymmetrical structures and they also allow us to extend considerably the range of impedances, even of symmetrical structures, which can be realized in the ladder form. The complex m structures, on the other hand, are chiefly of interest for their novel phase and attenuation characteristics. The novel phase characteristics are particularly important since the elementary constituents of the linear phase shift filters described in the accompanying paper usually turn out to be complex m sections.

The second general result is an increase in our knowledge of the relations existing between filters of different physical configurations. This last point is particularly important because it allows us to convert filters from one type of configuration to an equivalent type which may be better suited for purposes of practical construction. For example, it allows us to convert the general lattice filter, which is very convenient for theoretical purposes but is very difficult to build practically, into a composite of simple lattices and ordinary ladder sections.

Ideal Wave Filters *

By H. W. BODE and R. L. DIETZOLD

The increasing usefulness of wave filters in the telephone plant, together with rising standards of quality, emphasizes the need of a systematic method for approximating ideal characteristics as closely as we please. By an ideal filter is meant a network having the properties of a distortionless transducer over a given frequency range and suppressing all other frequencies. A design method is presented whereby an arbitrarily close approximation to these properties may be realized in a physical network. Examples of actual designs illustrate the engineering features involved in the practical application of the theory.

Introduction

IN the phenomenal advance of telephone practice during the past twenty years, almost every step has further restricted the distortion which individual parts of a transmission system can be allowed to introduce into the signal. The extension of circuits to great distances made it necessary that each link pass on to the next a more faithful copy of the signal so that the accumulated effects of many links might not endanger the intelligibility. The extension of telephone circuits to new uses, such as the transmission of pictures and the distribution of broadcast programs, imposed new demands for accuracy. Each of these has required rising standards of performance for wave filters. More than anything else, however, it has been the introduction of carrier methods, with their comparatively large utilization of selective structures, which has given prominence to the problem of reducing the distortion from wave filters. With the increase in length and complexity of carrier systems, the problem of providing wave filters which will have no harmful effect upon transmission has become one of increasing importance.

What this requires of the filters quickly appears if we recall that a structure which transmits *all* signals without distortion must (1) possess a characteristic impedance which is a pure resistance independent of frequency; (2) attenuate steady sinusoidal signals equally at all values of frequency; and (3) introduce a rotation in phase proportional to the frequency. In filter theory we need consider these requirements over only a limited band, since the signals which filters

* The reader is referred to the preceding paper entitled "A General Theory of Electric Wave Filters."

are meant to transmit, whether voice, telegraph, or television, are of a specified type having energies concentrated in certain portions of the frequency spectrum. We can therefore say that an ideal filter is one which has the ideal phase, impedance and attenuation properties in the frequency range of the desired signal and which totally suppresses all other frequencies.

The conventional ladder type filter structures which have been so extensively studied may be made to yield any desired suppression at the unwanted frequencies. In the range of wanted frequencies, however, they show wide departures from all three ideal properties. The impedance characteristic can be greatly improved by suitable elaboration of the filter structure itself, but to approximate uniformity of loss or linearity of phase shift it has been necessary to make use of supplementary networks of empirical design.[1]

The design of such corrective networks is by no means an easy task, primarily because the filter characteristics for which they are supposed to compensate change very rapidly with frequency in certain intervals. Nevertheless, much has been achieved. Thus it has been found possible to limit reflection coefficients to 2 per cent, in contrast with coefficients of 50 per cent not uncommonly tolerated in the systems of ten years ago. Improvements in the other characteristics have been comparable. In modern systems variations in attenuation of a few hundredths of a decibel, or in phase slope of a few per cent, can be attained if need be. These limits, however, demand the most patient and skillful design, and can seldom be met unless control of a single one of the characteristics is especially important. Since amplitude equalizers introduce non-linear phase, phase correctors non-uniform loss, and so on, the problem becomes increasingly difficult when close requirements must be met in several characteristics simultaneously.

By contrast, the method proposed in this paper gives the various characteristics simultaneously in a single network without recourse to auxiliary corrective structures. The method is a systematic one, requiring comparatively little in the way of cut and try design work. At the same time it preserves a measure of the flexibility of the existing technique, so that when considerable deviation from the ideal is tolerable in one or more characteristics, a corresponding economy of materials may be effected.

[1] The distortion problem has been discussed by several writers in this *Journal*. See, for example, S. P. Mead, "Phase Distortion and Phase Distortion Correction," April, 1928, p. 195; O. J. Zobel, "Distortion Correction in Electrical Circuits . . .," July, 1928, p. 438; C. E. Lane, "Phase Distortion in Telephone Apparatus," July, 1930, p. 493; E. B. Payne, "Impedance Correction of Wave Filters," October, 1930, p. 770.

The discussion which follows has a two-fold objective. The first is purely theoretical: to demonstrate that no matter how close the limits of deviation from the ideal may be set, there is a finite physical network all of whose characteristics meet these limits, except within a certain "transition interval" about each cut-off, which transition interval may also be taken as narrow as we please. This is by no means trivial; for it is known that no network, finite or infinite, can meet the ideal characteristics *exactly*.[2]

The second object is to guide the selection, from among the many networks which would meet the requirements of a given practical problem, of that one which meets them most economically. This part of the paper contains a number of examples, among them some which illustrate the use of slight empirical variations as a means of obtaining the highest measure of economy when wide deviations from the ideal are more tolerable in one respect than in others. The final example, which is segregated as Part III, deals with a situation met in picture transmission circuits, where the selectivity required is frequently small, but the effects of phase distortion may be very serious. Here a modification of the design technique leads to a filter which has comparatively modest selectivity but which exhibits a linear phase characteristic not only in the transmitting band but also in the range of rising attenuation.

Part I—Theoretical Analysis

Since linear phase shift is not available from ladder networks, the analysis will be based upon the more flexible lattice configuration. Although the lattice lends itself particularly well to the theoretical design problem, it is not so satisfactory for purposes of physical construction. After the paper design has been made, therefore, it will usually be desirable to convert it to a more suitable practical configuration. This can be done by methods described elsewhere.[3]

We may greatly simplify the theoretical discussion by ignoring the effects of parasitic dissipation—a simplification warranted by Mayer's Theorem,[4] which states that the attenuation resulting from dissipation

[2] This proposition is due to Dr. T. C. Fry, who showed that in a transducer possessing the steady-state characteristics of an ideal filter, a signal would arrive at the receiving terminals before it began to be impressed on the sending terminals. As this is absurd, we must conclude that no such system exists.

[3] H. W. Bode, "A General Theory of Electric Wave Filters," M.I.T. *Journal of Mathematics and Physics*, November, 1934. A summary of this article appears in this issue of the *Bell System Technical Journal*.

[4] H. F. Mayer, "Über die Dämpfung von Siebketten im Durchlässigkeitsbereich," *E. N. T.*, October, 1925, p. 335. His results were later somewhat extended by Feige and Holtzapfel, "Dämpfung und Winkelmass von Vierpolen mit geringen Verlusten," *T. F. T.*, July, 1932, p. 179. Even these latter results are capable of considerable generalization, so as to include other characteristics of the network besides the transfer constant.

is proportional to the derivative of the phase characteristic. The realization of a linear phase shift in the transmission band therefore automatically carries with it the satisfaction of the requirement of uniform loss in this range.[5] It can also be shown that the other characteristics of the network will not be appreciably affected by slight uniform dissipation.

Moreover, it is well known that the image impedance and transfer constant of a lattice structure are controlled by independent parameters.[6] We can, therefore, dissociate the problem of providing the required constant image impedance in the transmission band from that of providing the required loss and phase characteristics.[7] For the moment we shall fix our attention on the transfer constant.

With these simplifications our problem reduces to that of constructing a filter whose transfer constant on a non-dissipative basis represents a linear phase shift in the transmission band and an infinite loss in the attenuation bands, these being separated by narrow "transition intervals" in the neighborhood of the cut-offs. These transition intervals may be taken small at pleasure, but must be assigned in advance to insure the physical realizability of the network.

Formulation of Requirements—Low-Pass Filters [8]

If the impedances of the arms of a lattice are Z_x and Z_y, Fig. 1, it is well known that the image transfer constant and the image impedance are given by the expressions [9]

$$\tanh \frac{\theta}{2} = \sqrt{\frac{Z_x}{Z_y}}, \tag{1}$$

$$Z_I = \sqrt{Z_x Z_y}. \tag{2}$$

[5] Strictly speaking, a slight qualification should be placed upon this statement. Our process of approximating the ideal characteristics will lead to a phase shift which ripples about the desired linear characteristic, the number of ripples depending upon the number of elements used. As the number of elements is increased indefinitely, the linear characteristic is approximated more and more closely, but it is evidently not a necessary consequence of this that the slope of the ripples should approach constancy. We shall be able to show, however, that with the actual process used, the amplitude of the ripples decreases so rapidly that $dB/d\omega$ approaches constancy as B approaches linearity.

[6] This follows at once from equations (4) and (5), p. 220.

[7] A method of choosing the lattice parameters to give a substantially constant impedance in the transmission band has in fact already been obtained by W. Cauer, "Siebschaltungen," V. D. I. Verlag, Berlin, 1931; or "Ein Interpolationsproblem mit Funktionen mit Positivem Realteil," *Math. Zeit.*, November, 1933, p. 1. An alternative method will eventually be developed as a by-product of the present analysis.

[8] The extension to filters of other types is given on p. 225.

[9] G. A. Campbell, "Physical Theory of the Electric Wave Filter," this *Journal*, Vol. I, No. 2, November, 1922, p. 1.

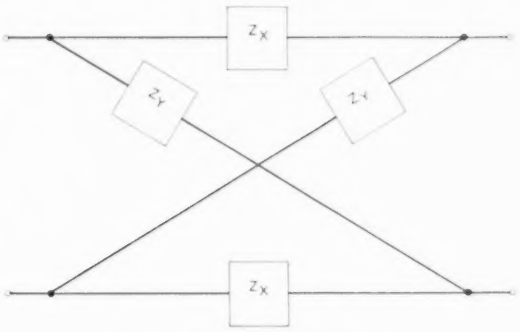


Fig. 1— The symmetrical lattice.

The relation (1) requires for transmission, i.e., for θ imaginary, that Z_x and Z_y differ in sign; for attenuation, i.e., for θ real, that Z_x and Z_y be alike in sign. In the case of the low-pass filter, this amounts to requiring correspondence of zeros (resonances) in one arm to poles (anti-resonances) in the other for $f < f_c$, and of zeros to zeros and poles to poles for $f > f_c$, where f_c, the cut-off, is a critical frequency which appears in one arm only.[10] If we denote these critical frequencies by $f_1, f_2, \cdots, f_r$ in the range below f_c, and by $f_1', f_2', \cdots, f_s'$ in the range above f_c, and if we make use of a well-known theorem [11] we readily find that Z_x and Z_y have forms similar to [12]

$$\left.\begin{aligned} Z_x &= iK_x f \frac{a_2 a_4 \cdots a_{r-1} a_r a_2' \cdots a_{s-1}'}{a_1 a_3 \cdots a_r a_1' \cdots a_s'}, \\ Z_y &= -\frac{iK_y}{f} \frac{a_1 a_3 \cdots a_r a_2' \cdots a_{s-1}'}{a_2 a_4 \cdots a_{r-1} a_1' \cdots a_s'}, \end{aligned}\right\} \tag{3}$$

[10] In the basic theory given by Dr. Campbell, in the paper just referred to, it is shown that in general a lattice having many natural frequencies is a milti-band-pass filter. The extension of the theory in the manner shown above, in which separate parameters for the control of the transfer constant and image impedance are obtained by imposing special conditions on the natural frequencies, thus rendering many bands confluent, was discovered and exploited independently by W. Cauer and one of the present writers (see Cauer, "Siebschaltungen" and later papers; or H. W. Bode, U. S. Patent No. 1828454, also "A General Theory of Electric Wave Filters," loc. cit.). The published work by Dr. Cauer gives a particularly complete discussion. It appears from a recent informal communication from Dr. Campbell to the authors, however, that this extension was also considered by him and was described briefly in the Yale-Harvard Lectures on Wave Filters delivered in 1923. The lectures have unfortunately not been published. Their content, however, is similar to that given in the discussion above.

[11] R. M. Foster, "A Reactance Theorem," this *Journal*, Vol. III, No. 2, April, 1924, p. 259.

[12] The cut-off factor a_c may appear in either the numerator or denominator of either form, and the line- and cross-arms may be interchanged. Otherwise the expression is general.

and hence that [13]

$$\tanh \frac{\theta}{2} = i \sqrt{\frac{K_x}{K_y}} f \frac{a_2 a_4 \cdots a_{r-1}}{a_1 a_3 \cdots a_r} \sqrt{a_c}, \tag{4}$$

$$Z_I = \sqrt{K_x K_y} \sqrt{a_c} \frac{a_2' \cdots a'_{s-1}}{a_1' \cdots a_s'}, \tag{5}$$

where the a's are shorthand notation for

$$a_j = 1 - \frac{f^2}{f_j^2}, \qquad a_j' = 1 - \frac{f^2}{f_j'^2}. \tag{6}$$

We shall have frequent occasion to distinguish between those critical frequencies of (4) which lie in the practical transmission band and those which lie in the transition interval. To this end we shall denote that interval by (f_A, f_B), where obviously $f_A < f_c < f_B$, and shall write P for the group of factors

$$P = i \sqrt{\frac{K_x}{K_y}} f \frac{a_2 a_4 \cdots a_{A-1}}{a_1 a_3 \cdots a_A} \tag{7}$$

which lie in the wanted band, and Q for the remainder

$$Q = \frac{a_{A+1} \cdots a_{r-1}}{a_{A+2} \cdots a_r} \sqrt{a_c} \tag{8}$$

which lie in the transition band. Then, obviously, (4) becomes

$$\tanh \frac{\theta}{2} = PQ. \tag{9}$$

Requirements on Transition Factors

With these formulæ before us, we are now prepared to attack the problem of meeting the double requirement of linear phase shift in the practical transmission band and infinite loss in the practical attenuation band. Expressed analytically, these requirements are simply

$$\tanh \frac{\theta}{2} = \tanh i \frac{\pi f}{2\alpha} = i \tan \frac{\pi f}{2\alpha}, \qquad f < f_A, \tag{10}$$

$$\tanh \frac{\theta}{2} = 1, \qquad f > f_B, \tag{11}$$

[13] It should be noted that, except for the cut-off factor a_c, (4) and (5) are entirely independent. That is, the frequencies $f_1, \ldots f_r$ may be chosen as we desire, in order to control the transfer constant, without in any way affecting the image impedance; and $f_1', \ldots, f_s'$ can be chosen at will without affecting θ. Similarly the constants $\sqrt{K_x/K_y}$ and $\sqrt{K_x K_y}$ may be chosen at will.

where α is a constant which determines the slope of the phase curve.

But it is well known that

$$i \tan \frac{\pi f}{2\alpha} = \frac{i\pi f}{2\alpha} \frac{\left(1 - \frac{f^2}{2^2\alpha^2}\right)\left(1 - \frac{f^2}{4^2\alpha^2}\right)\cdots}{\left(1 - \frac{f^2}{\alpha^2}\right)\left(1 - \frac{f^2}{3^2\alpha^2}\right)\cdots}. \tag{12}$$

If, then, in (4) we choose

$$\sqrt{K_x/K_y} = \pi/2\alpha,$$
$$f_1 = \alpha, \quad f_2 = 2\alpha, \quad \cdots, \quad f_A = A\alpha,$$

so that P becomes identical with the first A terms of (12), and if in addition we choose our unit of frequency so that [14] $(A + 1)\alpha = 1$, we readily see that in the transmitted range Q must equal

$$Q = \frac{(1 - f^2)\left(1 - \frac{f^2}{(1 + 2\alpha)^2}\right)\cdots}{\left(1 - \frac{f^2}{(1 + \alpha)^2}\right)\left(1 - \frac{f^2}{(1 + 3\alpha)^2}\right)\cdots}, \qquad f < f_A, \tag{13}$$

while by (9) and (11) in the attenuated range it must be given by

$$\frac{1}{Q} = P = \frac{i\pi f}{2\alpha} \frac{\left(1 - \frac{f^2}{2^2\alpha^2}\right)\cdots\left(1 - \frac{f^2}{(A-1)^2\alpha^2}\right)}{\left(1 - \frac{f^2}{\alpha^2}\right)\cdots\left(1 - \frac{f^2}{A^2\alpha^2}\right)}, \qquad f > f_B. \tag{14}$$

Expressed in terms of Gamma functions, (13) and (14) become [15]

$$Q = \frac{\Gamma^4\left(\frac{1}{2\alpha}\right)\Gamma\left(\frac{1-f}{\alpha}\right)\Gamma\left(\frac{1+f}{\alpha}\right)}{\Gamma^2\left(\frac{1}{\alpha}\right)\Gamma^2\left(\frac{1-f}{2\alpha}\right)\Gamma^2\left(\frac{1+f}{2\alpha}\right)}, \qquad f < f_A; \tag{15}$$

and

$$Q = \frac{1}{4\pi i}\frac{1-f}{2\alpha} \frac{\Gamma^4\left(\frac{1}{2\alpha}\right)\Gamma^2\left(\frac{f-1}{2\alpha}\right)\Gamma\left(\frac{f+1}{\alpha}\right)}{\Gamma^2\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{f-1}{\alpha}\right)\Gamma^2\left(\frac{f+1}{2\alpha}\right)}, \qquad f > f_B. \tag{16}$$

[14] This means that we express all frequencies in terms of the first critical frequency of (12) which falls in the transition interval.

[15] The necessary transformations may be found in Whittaker & Watson, "Modern Analysis," 12.13, 12.15, and 12.33.

Asymptotic Series for P and Q

If we now take the logarithm of (15), and apply Stirling's formula, we obtain the asymptotic series

$$\log Q = \log \frac{4^{1/\alpha}\Gamma^4(1/2\alpha)}{8\pi\alpha\Gamma^2(1/\alpha)} + \frac{1}{2}\log (1 - f^2) + \sum_{r=1}^{\infty} \frac{(-1)^r(2^{2r} - 1)B_r\alpha^{2r-1}}{2r(2r-1)} \left[\frac{1}{(1+f)^{2r-1}} + \frac{1}{(1-f)^{2r-1}} \right], \quad (17)$$

where the B's are Bernoulli numbers. Since Stirling's formula holds only for $z > 0$, this expansion is valid, as inspection of equation (15) will show, only for $f < 1$, but as the unit of frequency was so chosen that $f_A < 1 < f_B$, this includes the entire wanted band, and none of the attenuating range.

If we apply a similar process to (16) we are again led to (17), except that now the range of validity is $f > 1$. But this includes the entire attenuating range, and none of the wanted band.

That is, the single formula (17) represents the lacunar [16] function Q in both ranges in which it is well defined.

We shall now determine the transition factors $a_{A+1}, a_{A+2}, \cdots, a_c$ by comparison of (8) with (17). If we adopt the notation

$$f_{A+1} = 1 + c_1, \quad f_{A+2} = 1 + c_2, \quad \cdots, \quad f_c = 1 + c_m, \quad (18)$$

so that the c's measure, not the critical frequencies themselves, but their displacement from unit frequency, each factor of (8) has the characteristic form

$$1 - \frac{f^2}{(1+c_j)^2} \equiv \frac{1 - f^2}{(1+c_j)^2}\left(1 + \frac{c_j}{1-f}\right)\left(1 + \frac{c_j}{1+f}\right);$$

whence (8) becomes

$$Q = K \frac{\left(1 + \frac{c_1}{1-f}\right)\left(1 + \frac{c_1}{1+f}\right)\cdots\sqrt{1 + \frac{c_m}{1-f}}\sqrt{1 + \frac{c_m}{1+f}}}{\left(1 + \frac{c_2}{1-f}\right)\left(1 + \frac{c_2}{1+f}\right)\cdots\left(1 + \frac{c_{m-1}}{1-f}\right)\left(1 + \frac{c_{m-1}}{1+f}\right)} \sqrt{1 - f^2}, \quad (19)$$

where K is a constant multiplier which depends on the c's. We will neglect it in this analysis since it may be readily determined later from the condition that $Q = 1$ when $f = 0$.

[16] A lacunar function is one which is well defined in several regions, but not capable of analytic continuation from one to the other.

The logarithm of Q is, of course, the sum of the logarithms of the individual factors of this expression. Expanding these as series of powers of $1/(1-f)$ or $1/(1+f)$, and collecting terms of like degree in $1/(1-f)$ and $1/(1+f)$, we obtain

$$\begin{aligned}\log Q = {} & \frac{1}{2}\log(1-f^2) \\ & + \left(c_1 - c_2 + c_3 - \cdots \pm \frac{1}{2}c_m\right)\left[\frac{1}{1-f} + \frac{1}{1+f}\right] \\ & - \frac{1}{2}\left(c_1^2 - c_2^2 + c_3^2 - \cdots \pm \frac{1}{2}c_m^2\right)\left[\frac{1}{(1-f)^2} + \frac{1}{(1+f)^2}\right] \\ & + \frac{1}{3}\left(c_1^3 - c_2^3 + c_3^3 - \cdots \pm \frac{1}{2}c_m^3\right)\left[\frac{1}{(1-f)^3} + \frac{1}{(1+f)^3}\right] \\ & - \cdots,\end{aligned} \tag{20}$$

where the sign of c_m is plus or minus according as m is odd or even.[17] As the terms of (20) are similar in form to those of (17),[18] we can make the first m terms identical. This leads to the equations

$$\left.\begin{aligned} c_1 - c_2 + c_3 - \cdots \pm \frac{1}{2}c_m &= -\frac{3}{2}B_1\alpha, \\ c_1^2 - c_2^2 + c_3^2 - \cdots \pm \frac{1}{2}c_m^2 &= 0, \\ c_1^3 - c_2^3 + c_3^3 - \cdots \pm \frac{1}{2}c_m^3 &= +\frac{15}{4}B_2\alpha^3, \\ c_1^4 - c_2^4 + c_3^4 - \cdots \pm \frac{1}{2}c_m^4 &= 0, \\ c_1^5 - c_2^5 + c_3^5 - \cdots \pm \frac{1}{2}c_m^5 &= -\frac{21}{2}B_3\alpha^5, \\ \cdots \quad \cdots \quad \cdots \quad &\cdots, \end{aligned}\right\} \tag{21}$$

whose simultaneous solution gives the desired transition factors.

The number m of transition factors used will depend upon the desired approximation to ideal characteristics in the practical transmitting and

[17] When the c's are evaluated it will appear that these series are all absolutely convergent—so that their termwise sum correctly represents $\log Q$—at all positive frequencies outside the interval $(1 - c_m, 1 + c_m)$. As the complexity of the network is increased, in the approach toward ideal characteristics, the interval of non-convergence closes on the reference frequency 1, and is contained by the given transition interval.

[18] The first term of (17) does not contain f, and may therefore be neglected for the same reason which led us to neglect K in (19).

attenuating ranges and the allowable width of the transition interval. It can best be determined by inspection of results given later.

The result of solving the equations (21) for the ratios c_j/α for values of m between 1 and 5 is given in the following table.

TABLE I
SPACING OF TRANSITION FACTORS

Number of Factors	c_1/α	c_2/α	c_3/α	c_4/α	c_5/α
1	−0.50000				
2	−0.14645	+0.20711			
3	−0.05032	+0.67731	+0.95526		
4	−0.01897	+0.86157	+1.49180	+1.72252	
5	−0.00760	+0.93809	+1.74806	+2.30277	+2.50080

The first of these solutions corresponds to a single frequency, the cut-off, in the transition interval. It follows the uniformly spaced critical frequencies of the practical transmission band at one-half the uniform spacing, α. The other solutions represent networks having, in addition to the cut-off, rational factors which vanish in the transition interval.

When these values of the c's are used in equation (8), with due regard for (6) and (18), the form of Q is completely determined. For example, the frequency pattern corresponding to the case $m = 3$, is illustrated by Fig. 2.

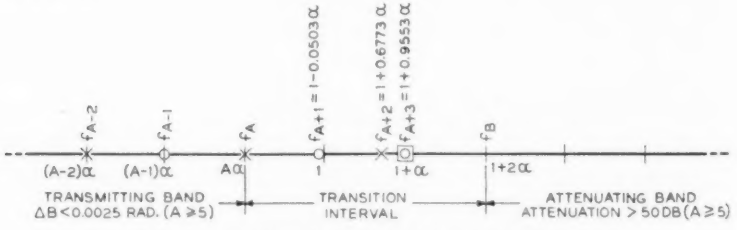


Fig. 2—Location of transition factors with $m = 3$.

Nature of the Approximation

How closely we approach ideal characteristics by this method depends on how nearly $\log Q$ is represented by the first m terms of (17) and (20). In both cases the series of omitted terms can be written in the form

$$A_{m+1}\alpha^{m+1}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right] + A_{m+2}\alpha^{m+2}\left[\frac{1}{(1-f)^{m+2}} + \frac{1}{(1+f)^{m+2}}\right] + \cdots, \quad (22)$$

where the A's are constants. In (20) this series is convergent. In (17) it is merely asymptotic. It is known, however, that the error due to ending (17) at any term is numerically less than the first omitted term. Since we are at present interested in small values of α, therefore, we can estimate the error in the approximation from the first term alone.

Inspection of this term shows that the error is greatest in the vicinity of the transition interval, where the factor $1/(1-f)$ is large. It depends upon all three of the quantities f, α and m; but by choosing them in the proper order there is no difficulty in showing that an indefinitely close approximation can be obtained.

The transition interval must first be selected on the absolute frequency scale. It may be as small as we choose. Next, a value of m must be chosen. What value is used is immaterial for our present purposes, although it is important for later applications. Finally, α must be taken small enough so that all transition factors lie in the prescribed transition interval. Otherwise it may be varied at will. But by choosing it small enough, the error of approximation (22) can obviously be reduced without limit for any value of f outside the interval (f_A, f_B). We may thus conclude that only considerations of expense and of manufacturing precision restrict the accuracy of approach to the ideal filter.

For purposes of future reference approximate formulæ for the attenuation and phase in the limiting condition are given below:

$$e^{-A} \doteq -\frac{1}{2} A_{m+1}\alpha^{m+1}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right], \tag{23}$$

$$B \doteq \frac{\pi f}{\alpha} + A_{m+1}\alpha^{m+1}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]\sin\frac{\pi f}{\alpha}. \tag{24}$$

It will be seen that the attenuation rises monotonically as we recede from the transmission band while the phase curve ripples about the ideal straight line in a sinusoid of varying envelope. The ripples, of course, increase in frequency as α is diminished but since the exponent $m+1$ is always at least 2 they flatten out so rapidly that $dB/d\omega$ approaches constancy nevertheless. We may also observe that, although the absolute time of delay increases indefinitely as α decreases, it varies only as $1/\alpha$, whereas the precision of approximation can be made indefinitely great by choosing m large.

Filters of Other Types

While the preceding analysis has been restricted formally to low-pass filters, its application to filters of other transmission types is a

simple matter. We need merely repeat, in each transmission or transition interval, the rules for frequency spacing we have already developed.

The method can be understood from the study of a linear phase shift high-pass filter. Since linear phase shift demands arithmetic spacing of critical frequencies in the transmission band, it is clear that the desired characteristic cannot be obtained over the complete transmission band of the high-pass filter with a *finite* network. This difficulty will, however, be ignored for the moment. A method of modifying the analysis to give a finite filter having a linear phase characteristic in a finite interval above the cut-off will be described later.

We begin, then, by assigning to the transmission band of the structure an infinite, evenly spaced chain of critical frequencies, as in (13). The group of transition factors must evidently simulate the reciprocal of this value in the attenuation range, if the condition of high loss is to be realized; while in the transmission band, they must simulate the P of our earlier analysis if we are to obtain a linear phase characteristic. These conditions would be met by using for our transition function the reciprocal of (19), using for the c's the same values as before. Such a group, however, is not physically realizable as part of a high-pass transfer constant, since the rational factors would occur outside the theoretical transmission band. If, however, we transfer the factor $(1 - f^2)$ from (13) to (14), and seek a new Q whose values will take the reciprocals of the old, thus altered, we obtain a series identical with (17) except for a change of sign in every term but $\frac{1}{2} \log (1 - f^2)$. This change, however, reverses the sign of the right-hand members of (21), and therefore changes the sign of each c. The new solution then is the same as the original solution except that the factors occur in reverse order on the frequency scale. They can thus appropriately be combined with the remaining portion of the high-pass transfer constant expression.

A linear phase shift band-pass filter can be constructed similarly. The groups of transition factors associated with the upper and lower cut-offs should follow the arrangements prescribed, respectively, for low-pass and high-pass filters. An illustration will be found in Part II.

The Impedance Property

It will be recalled that the problem of approximating the ideal transmission characteristics for each type of filter was solved only on the assumption that the image impedance could be adjusted to a nearly uniform value in the practical transmitting band. We can

now quickly show how the desired impedance is to be obtained. It is merely necessary to observe that for any filter there exists a complementary structure with the same arrangement of critical frequencies, but having the transmitting and attenuating bands interchanged. The complementary structure is found by replacing the Z_y branch of the original lattice by the inverse impedance $Z_y' = R^2/Z_y$. When these are substituted in (1) and (2), the new transfer constant, θ', is found to be

$$\tanh \frac{\theta'}{2} = \sqrt{\frac{Z_x}{Z_y'}} = \frac{1}{R}\sqrt{Z_x Z_y} = \frac{1}{R} Z_I,$$

and the new image impedance, Z_I',

$$Z_I' = \sqrt{Z_x Z_y'} = R\sqrt{\frac{Z_x}{Z_y}} = R \tanh \frac{\theta}{2}.$$

Thus, for any filter, the problem of adjusting the image impedance to the constant R in its transmitting band is the same as the problem of adjusting $\tanh \theta/2$ to 1 in the attenuating band of the complementary filter. The latter problem, however, is merely a restatement of our original requirement of high loss in attenuating bands and has already been studied for various types of filters.

It follows from this relation that the transfer constant expressions which are appropriate for low-pass and band-pass filters furnish suitable solutions for the impedance problem in high-pass and band-elimination structures. We might also use our high-pass transfer constant expression as a low-pass impedance characteristic except for the difficulty previously mentioned that it requires an infinite number of elements. This difficulty can be avoided, however, by observing that by interchanging coils and condensers we can convert any low-pass filter into a high-pass structure having the same characteristics on a reciprocal frequency scale. We can thus use the finite low-pass solutions to obtain the required finite high-pass filter having high attenuation. For example, if we begin with a low-pass filter having three evenly spaced critical frequencies and a half spaced cut-off the resulting critical frequencies (including the cut-off) are in the ratio $1 : 7/6 : 7/4 : 7/2$.

The device of inverting the frequency scale is, of course, not available to produce a finite high-pass filter having linear phase shift throughout its transmission band since the linear phase property is thereby destroyed. It can be used, however, to produce a finite filter having linear phase shift for a limited region above its cut-off.

To see this, it is merely necessary to observe that the set of rational factors appearing in the low-pass image impedance expression described in the preceding paragraph must approximate the reciprocal of the cut-off factor at lower frequencies. We can therefore use such a set of factors to replace the upper cut-off factor of a band-pass filter, obtaining thereby a high-pass structure which approximates the ideal characteristics over a portion of the transmitting band.

If the cut-off factor of the low-pass filter transfer constant be similarly replaced by rational factors, there results an all-pass "delay network" having a constant impedance and a phase characteristic linear below the original cut-off frequency. This network is of particular interest for its relation to the classic problem of the simulation of a smooth line. As it stands, the network evidently simulates an ideal dissipationless line. To include the effects of dissipation we need merely add resistance and leakance to the coils and condensers in the proportions in which they occur in the actual line.

Part II—Design of Practical Filters

Thus far we have been interested primarily in demonstrating that an indefinitely close approximation to the ideal characteristics could be obtained when all restrictions with respect to economy of elements were removed. In practical designs, on the other hand, we wish to approximate the ideal characteristics only within moderate limits, and our interest centers upon the choice of the most economical network which will prove satisfactory. We must now reappraise the theory from this point of view.

One question which must be examined is that of determining values for m and α which will result in the most economical network meeting a prescribed standard of performance. A second is concerned with the possibility of changing the nature of the approximation with respect either to the frequency, or to the relative emphasis laid upon the phase and attenuation characteristics. In many practical designs such changes can be obtained by slight modifications of the theoretical design parameters and lead to corresponding economies in the use of elements. In investigating both questions we must remember that since α is no longer necessarily small, as it was in the theoretical analysis, the frequency interval actually occupied by the transition factors may be appreciable. Consequently it becomes important to investigate the behavior of the network in this part of the frequency range with more care than was hitherto necessary.

The variety of possible design requirements precludes the possibility of a thorough analytic treatment of these questions. The choice of

the most economical network meeting given requirements consequently cannot always be made without trial. The procedure may, however, be considerably facilitated by a study of the curves and illustrative material given in the sections which follow. The first two sections show the quantitative relations to be expected when the theoretical design parameters are adhered to strictly. The remaining sections indicate modifications obtainable by making slight changes in the theoretical parameters.

Approximate Computation of Network Characteristics

When the frequency in which we are interested is not too close to the transition interval, an approximate determination of the phase and attenuation characteristic is most easily made from (23) and (24). The A's appearing in these expressions are shown in the accompanying table.[19] In addition to A_{m+1} the table also supplies values of A_{m+2}

TABLE II

COEFFICIENTS IN SERIES EXPANSIONS FOR APPROXIMATION ERRORS IN PHASE AND ATTENUATION CHARACTERISTICS

m	1	2	3	4	5
A_{m+1}	−0.063	−0.044	−0.051	−0.084	−0.17
A_{m+2}	−0.063	+0.00011	+0.10	+0.41	+1.52
A_{m+3}	−0.0078	+0.050	−0.047	−1.07	−7.60

and A_{m+3}, for use if additional terms in the general expression (22) are desired.

A study of equations (23) and (24) shows that, aside from the constant factor A_{m+1}, each expression can be resolved into two factors by means of which the contributions of the various design parameters can be somewhat segregated. The first factor, α^{m+1}, is chiefly important in determining the effect of various choices of α and m on the approximation error, while the factor $\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]$ expresses the variation of the network characteristics with frequency. In order to facilitate design work the quantity

$$-20 \log_{10} \frac{A_{m+1}}{2}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]$$

[19] In preparing the table, coefficients of corresponding terms in the series expansions for (17) and (20) have been combined, so that the coefficients as given represent the accumulated errors of both approximations.

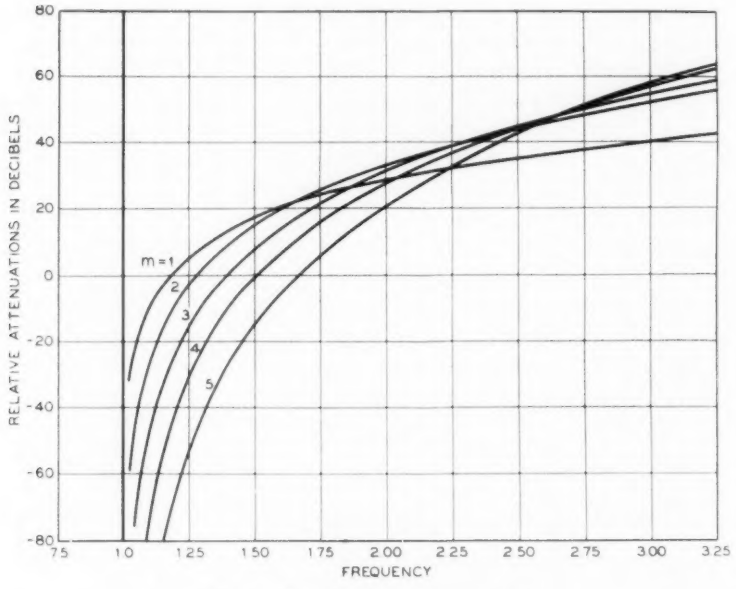


Fig. 3—Chart for loss computations.

has been computed for values of $f > 1$ and is shown plotted for various m's in Fig. 3. The approximate attenuation, in db, for any given values of α and m can be obtained from the chart by adding $20(m + 1) \log_{10} 1/\alpha$ to the appropriate curve.

A similar chart for the phase characteristic is furnished by Fig. 4, which represents the quantity $\frac{180A_{m+1}}{\pi}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]$. The values given are arithmetic, although the scale is logarithmic. The approximate envelope of the ripple in the phase characteristic about the ideal straight line can therefore be found, in degrees, by multiplying the chart values by α^{m+1}.

In using these charts it should be remembered that they are based upon the approximate formulæ (23) and (24) which fail in the vicinity of the transition interval. The results, therefore, should always be checked by an exact computation.[20] It should also be observed that in complicated filters the numerical departure of $\tanh \theta/2$ from its ideal value in most frequency ranges is very small. The effects of slight errors in calculations or of small deliberate variations in the

[20] See, for example, the comparisons in Figs. 11 and 12.

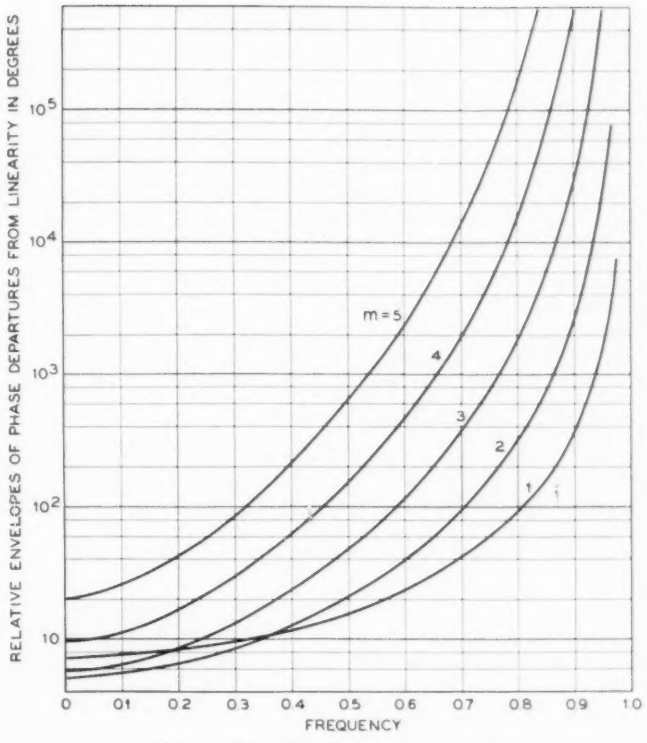


FIG. 4.—Chart for phase computations.

design parameters, may therefore be correspondingly important.[21] Since slight adjustments in the design parameters will normally occur, these charts are chiefly of value in making preliminary estimates.

It is apparent that the approximation error at a given frequency can be diminished either by increasing m or reducing α. Element for element, an increase in m is much the more powerful method. Since the total number of elements in the network is nearly proportional to

[21] A simple example is furnished by the choice of the numerical constant multiplying $\tanh \theta/2$ as a whole. It will be remembered that the constant was left undetermined in the solution for the c's. In the original equation (14) it was chosen to give the best characteristics in the neighborhood of $f = 0$. In preparing Figs. 3 and 4, on the other hand, it was chosen with reference to the characteristics near $f = \infty$, since the error expression used in these figures vanishes at that point. The two conditions are very nearly equivalent; as we can see, for the half-spaced cut-off solution at least, by means of Wallis' theorem. Since they are not identical, however, a change from one to the other may produce a relatively large, though practically unimportant, effect at extreme frequencies.

$m + (1/\alpha)$ it would therefore appear that the most economical structure meeting given requirements will be obtained by using a large m in combination with a large α. This procedure is, however, restricted by two considerations. The first is chiefly theoretical. Since the series we have been using is merely asymptotic, the successive terms obtained by choosing progressively higher m's eventually grow larger. For ordinary values of α, however, the value of m at which the series begins to diverge lies beyond the range of practical interest. A more important limitation is the fact that as we increase the number of transition factors, the width of the transition interval, as measured in terms of α, also increases. Thus, the spread between the last uniformly spaced critical frequency and the cut-off, which is $\alpha/2$ for $m = 1$ and about 2α for $m = 3$, has risen to more than 3.5α for $m = 5$. In each case a certain additional allowance is of course required for the region of rising attenuation beyond the cut-off. When the transition interval is fixed on an absolute frequency scale, therefore, the permissible values of m will depend upon the choice of α. Unless the transition interval is unusually broad only low values of m will be of practical interest.

Illustrative Characteristics

The curves shown in Figs. 3 and 4 are not of use in the neighborhood of the transition interval. To supplement them, therefore, exact computations on a number of typical structures have been made. One set was obtained by choosing $\alpha = 1/12$ and computing the characteristics corresponding to various m's. The resulting phase characteristics are shown by Fig. 5. Since the departures from linearity are too small to be noticeable when the characteristics as a whole are drawn, the figure shows only the departures themselves in terms of an envelope similar to that used for Fig. 4. The curves are drawn approximately as far as the last evenly spaced critical frequency which marks the practical limit of the range within which a high degree of phase linearity is to be expected. Since the curves vary rapidly in this vicinity, however, the fact that they are merely envelopes is important in determining the exact performance of the structure. Curves of the phase characteristics in the transition interval will be given later.

The attenuation characteristics are shown by Fig. 6. As m is increased, the cut-off moves to successively higher frequencies because of the progressively broader intervals consumed by the transition factors. Once past the cut-off, however, the curves for large values of m rise more rapidly and quickly cross the others.

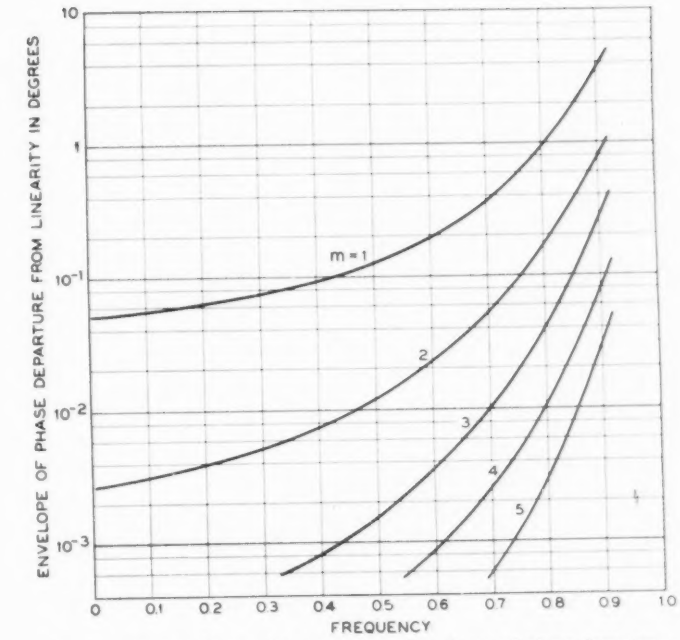


Fig. 5—Low pass filters with $\alpha = 1/12$. Envelopes of phase departures.

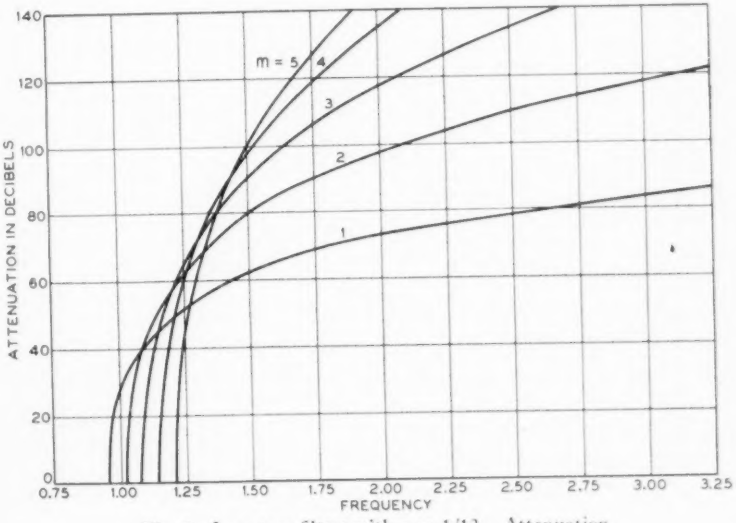


Fig. 6—Low pass filters with $\alpha = 1/12$. Attenuation.

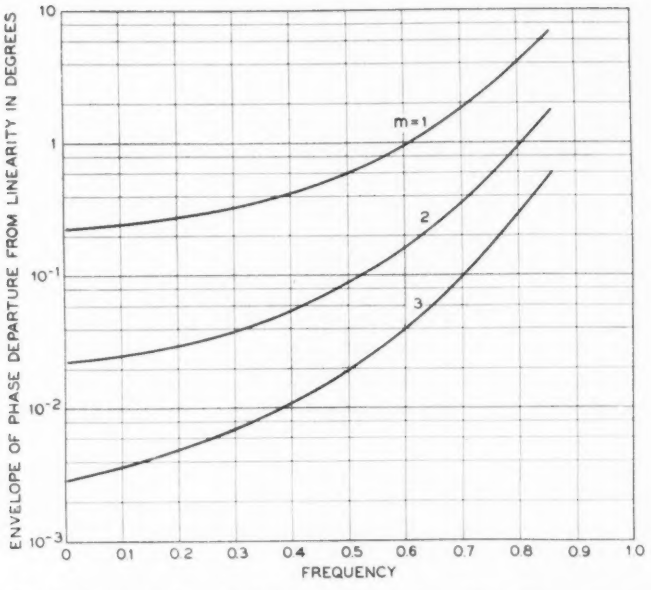


Fig. 7—Low pass filters with $\alpha = 1/6$. Envelopes of phase departures.

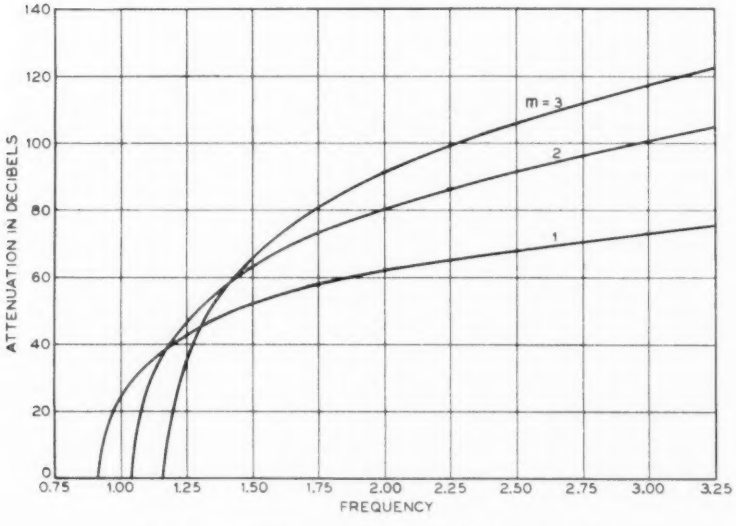


Fig. 8—Low pass filters with $\alpha = 1/6$. Attenuation.

A second set of characteristics was obtained by choosing $\alpha = 1/6$ and adding various groups of transition factors in a similar fashion. The results are shown by Figs. 7 and 8. The characteristics are drawn only for m's between 1 and 3 in this case, since with larger m's the

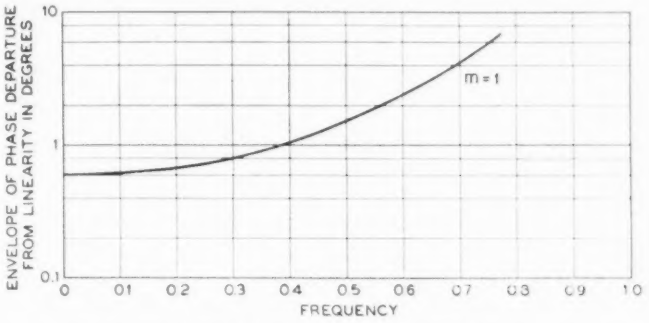


Fig. 9—Low pass filter with $\alpha = 1/4$. Envelope of phase departures.

transition interval becomes disproportionately wide in comparison with the practical transmission range. Still a third set, corresponding to $\alpha = 1/4$ and $m = 1$ is shown by Figs. 9 and 10.

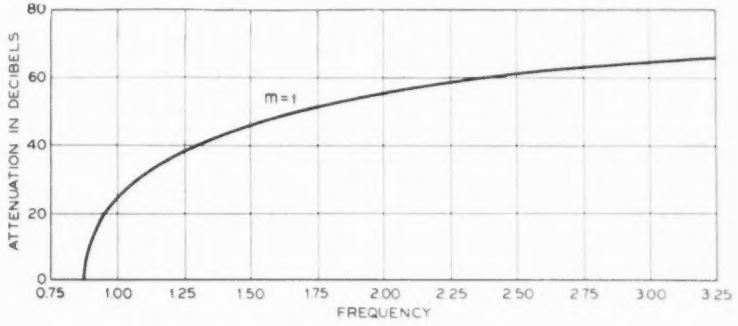


Fig. 10—Low pass filter with $\alpha = 1/4$. Attenuation.

As an illustration of the accuracy to be expected from the approximate method, a comparison between the results obtained by this method and the exact characteristics is shown in Figs. 11 and 12 for the cases $m = 1$ and $m = 2$ of Figs. 5 and 6. On the logarithmic scales used for the figures, the curves appear to be in good agreement almost up to the transition interval. The actual numerical departures in the vicinity of that interval, however, are quite large.

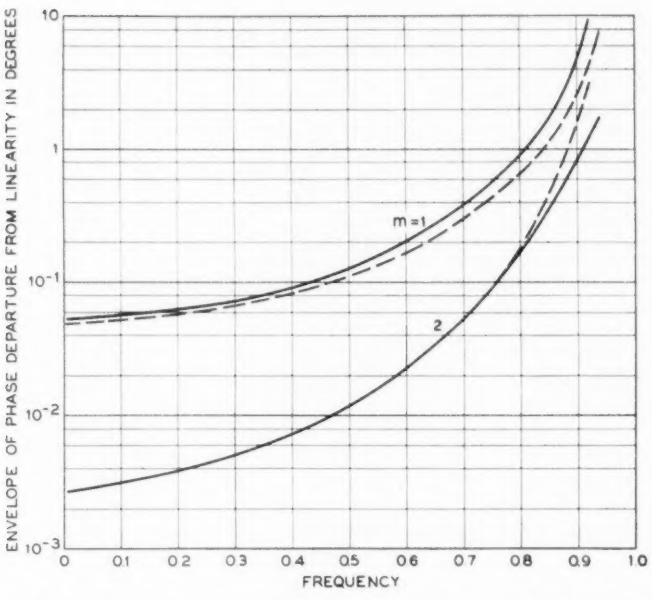


Fig. 11—Approximate and exact envelopes of phase departure for low pass filters with $\alpha = 1/12$.

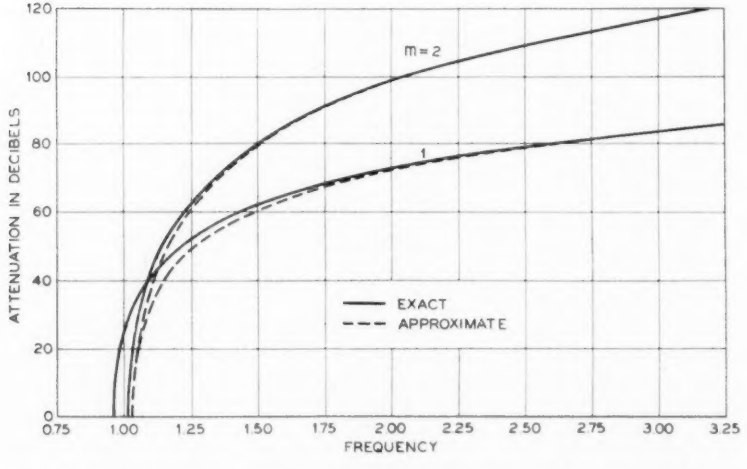


Fig. 12—Approximate and exact attenuation characteristics for low pass filters with $\alpha = 1/12$.

Image Impedance Characteristics

In virtue of the relationship previously developed between the image impedance of a given filter and the transfer constant of its complement, the curves just given might also be used to determine the impedance characteristic. However, the precision required in the approximation of Z_I to R in practical filter design is much less than that required in the approximation of $\tanh \theta/2$ to unity. A satisfactory characteristic can therefore be obtained with a much smaller number of critical frequencies. In a low-pass filter, for example, one or two impedance controlling frequencies is usually sufficient. With such a small number of critical frequencies the analytical machinery we have set up is unnecessarily cumbersome. The problem can be solved more effectively by simple cut and try methods, or by the methods advanced by Cauer [7] and Zobel.[22] For the sake of completeness, however, several illustrative characteristics are given in Fig. 13. They correspond to the choice of impedance controlling frequencies

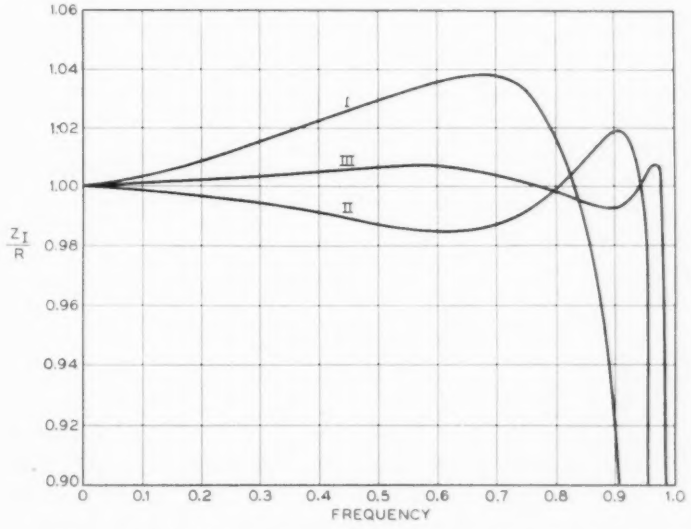


Fig. 13—Typical low pass filter image impedance characteristics.

[7] "Siebschaltungen," loc. cit.

[22] This *Journal*, Apr. 1931, p. 284. Zobel's work is not stated in terms of the lattice parameters. A simple *m*-type termination (of low-pass or high-pass type) can be identified with a lattice image impedance having one impedance controlling frequency while an *mm'*-type termination can be identified with a lattice impedance having two such frequencies. The numerical values he gives can therefore readily be adapted to the lattice design problem.

shown in Table III. An illustration of the results obtainable with the present method using a large number of critical frequencies, is furnished by Fig. 14, since the curve can evidently be interpreted as a representation of Z_I/R for a certain high-pass filter.

TABLE III

IMPEDANCE CONTROLLING FREQUENCIES CORRESPONDING TO CHARACTERISTICS OF FIG. 13

I	II	III
$1.250\,f_c$	$1.048\,f_c$ $1.448\,f_c$	$1.013\,f_c$ $1.096\,f_c$ $1.584\,f_c$

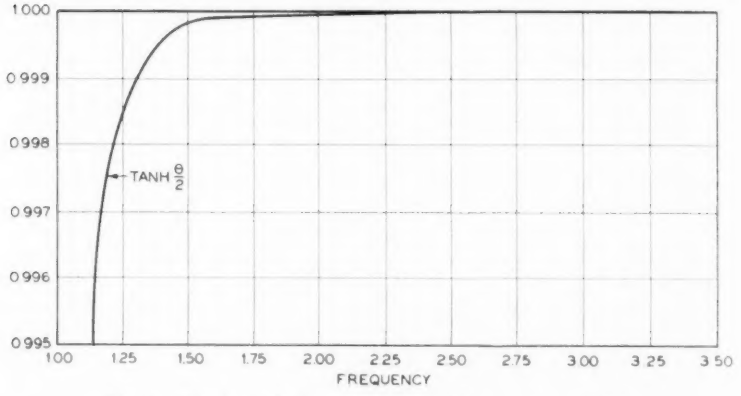


Fig. 14—Tanh $\theta/2$ for low pass filter with $m = 2$, $\alpha = 1/12$.

Weighting the Approximation

In the limiting case in which α is very small while m is fixed, the methods we have followed give the best obtainable results with respect to both attenuation and phase, for the errors in both characteristics depend upon higher order terms and become negligible as α approaches zero. In a practical design, for which α is finite, on the other hand, it will frequently be desirable to make slight adjustments in such parameters as the transition factors or the constant multiplier of the tanh $\theta/2$ expression in order to take some account of the higher order terms. The effect may be either to improve both the phase and attenuation characteristics or, more usually, to improve one at the expense of the other.

The general nature of the problem is illustrated by Fig. 14, which

represents a sketch of tanh $\theta/2$ corresponding to the $m = 2$ curve of Fig. 6. It will be seen that the curve rises monotonically toward the line unity, at which $\theta = \infty$. What we should evidently like to obtain by slight alterations in the design parameters is a curve which rises more rapidly, or perhaps one which ripples about unity. It is also evident that the curve approximates unity so closely that even slight adjustments may produce a radical effect. To take the simplest possibility, if the constant multiplier of tanh $\theta/2$ is slightly increased, so that the curve crosses unity at a finite frequency, the appearance of the resulting attenuation characteristic will be greatly altered. The net gain in the general level of attenuation secured, however, will be not more than 6 db. Similar remarks might be made with respect to the phase characteristic.

The relation between the phase and attenuation characteristics where such adjustments are made can be illustrated most easily by reference to the elementary half-spaced cut-off solution for the transition factors. It will be recalled that this solution was obtained by equating the coefficients of the first powers of $1/(1 - f)$ and $1/(1 + f)$ in (17) and (20). The approximation error thus depends chiefly upon the succeeding term involving $1/(1 - f)$ and $1/(1 + f)$ to the second power. A study of the expression shows that the error makes Q too small in both the transmitting and attenuating ranges. If the phase characteristic is the more important this error can be partly compensated by slightly increasing the normal half-space between the cut-off and the preceding critical frequency. On the other hand, the attenuation will be improved if the interval between the cut-off and the preceding critical frequency is decreased. To a more limited extent, both characteristics can be improved by increasing the constant factor which multiplies tanh $\theta/2$ as a whole.

A similar study might be made of the other groups of transition factors, although the discussion would naturally become more complicated. In general it appears, as with the half-spaced cut-off solution, that the attenuation characteristic will be improved by a slight decrease in the spacings of the transition factors, while the phase characteristic will be improved if they are slightly increased. It should be remarked, however, that as the network becomes more complicated, either by a reduction in α or an increase in m, the desirable modifications in the theoretical spacings are reduced. This becomes evident if it is recalled that the transition factor spacings are proportional to α while the error is roughly proportional to α^{m+1}. It is therefore to be expected that the appropriate modifications in the spacings between transition frequencies will be of the order of magnitude of α^m times their original values.

The relationship between the phase and attenuation characteristics can be seen in another light if we observe that the improvement in attenuation which comes from the use of several transition factors is due essentially to a progressive decrease in the interval between critical frequencies as the cut-off is approached. In the final solution, for example, the intervals between critical frequencies are initially almost equal to the constant interval α. Thus in this solution, the interval between f_A and f_{A+1} is 0.992α and that between f_{A+1} and f_{A+2} is 0.945α. As the cut-off is approached, however, the interval gradually decreases to about 0.2α. In the transition interval, consequently, the phase characteristic is originally almost linear and curves upward sharply near the cut-off. Thus if the phase requirement is not severe we can consider that the first part of the transition region falls within the practical transmitting band, thereby securing a better attenuation characteristic than would be possible if the spacing of critical frequencies in the transmitting band were strictly uniform. A sketch of the phase characteristics through the transition interval

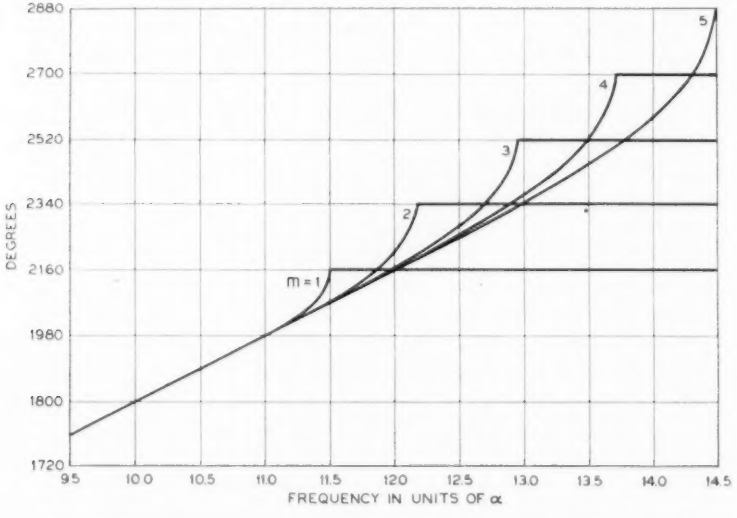


Fig. 15—Transfer constant phase shift in the transition interval; $\alpha = 1/12$.

for the networks corresponding to Figs. 5 and 6 is shown by Fig. 15. The last evenly spaced critical frequency falls at 11α.

In the extreme case when no phase requirement is imposed, it is reasonable to expect that the best attenuation characteristic will be

obtained if the progressive reduction in the spacing between critical frequencies extends over the complete transmission band, so that the phase characteristic should resemble that of familiar ladder type filter structures by becoming continually steeper as the cut-off is approached. The exact arrangement will, however, depend upon the desired type of best approximation to perfect suppression. If the approximation is to be best at frequencies most remote from the cut-off, the critical frequencies must be evenly spaced along an ordinary arc sine curve. In the Tschebycheffian type of approximations studied by Cauer, on the other hand, the spacing must be uniform along the arc of a certain sn function.

Design of a Band-Pass Filter

To illustrate the manner in which small modifications of the theoretical frequency spacings may be employed to control the relative emphasis placed on the phase and attenuation characteristics, we may consider the design of a practical band-pass filter. Suppose that the practical transmitting band is the 2,250-cycle interval between 11,375 and 13,625 c.p.s., in which the approximation of the phase characteristic to linearity is specified by the requirement that $\partial B/\partial\omega$, the so-called "delay," deviate from its average value by less than 0.1 millisecond. The transition intervals are 500 c.p.s. each, beyond which the loss is to be not less than 50 db.

The comparatively liberal tolerances suggest that the approximation furnished by $m = 1$ will be adequate. We notice that we can fit 10 uniform intervals of 250 c.p.s. between 11,250 and 13,750 c.p.s., which locates the half-spaced cut-offs at 11,125 and 13,875 c.p.s. respectively. When the characteristics corresponding to this design are checked, it is found that the phase characteristic is rather better than required, while the loss characteristic is weak.

We then turn to the solution with $m = 2$, making a compensating reduction in the number of uniform intervals. The critical frequency allocation for this case is shown in Table IV. This arrangement meets

TABLE IV

CRITICAL FREQUENCY ALLOCATION FOR LINEAR PHASE SHIFT BAND-PASS FILTER

$m = 1$	$m = 2$	Modified	Limits of Required Linear Region
11,125	11,198	11,174	
11,250	11,289	11,265	11,375
11,500	11,500	11,500	
. . . .	. . . .	. . . .	
13,500	13,500	13,500	13,625
13,750	13,711	13,735	
13,875	13,802	13,826	

the loss requirement with a large margin of safety, but the phase shift curve departs seriously from linearity near the last useful frequencies, which fall in the shortened intervals of the transition factors.

With these two attempts as guides, a compromise frequency pattern which exactly suits the conditions of the problem is readily arrived at. In contrast to the transition factor spacings of 0.853α and 0.353α, as shown by the solution for $m = 2$, those actually adopted are 0.94α and 0.375α so as to make the first transition spacing more nearly uniform with those in the pass-band. The indicated frequency pattern is shown in the third column of the table.

As the values of these transition factors near the band edges are somewhat too large they lead to larger undulations of the phase characteristic in those regions than near the band center. The approximations can be rendered more uniform throughout the band without serious consequence to the loss characteristic by multiplying the tangent expression by a constant slightly smaller than unity. In this case the value chosen was $K_1 = 0.9975$.

The final "delay" and loss characteristics, corrected for the effects of dissipation, are exhibited by Figs. 16 and 17. A noteworthy result

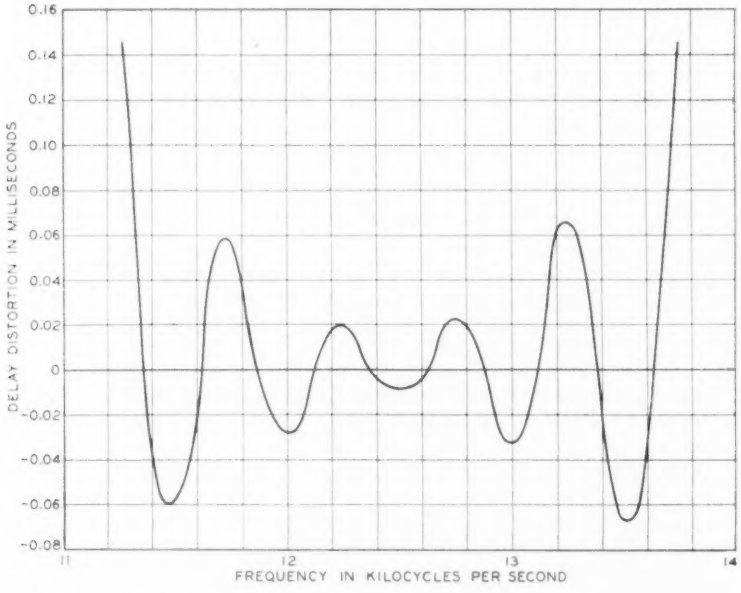


Fig. 16—Variation in the phase slope of a band pass filter.

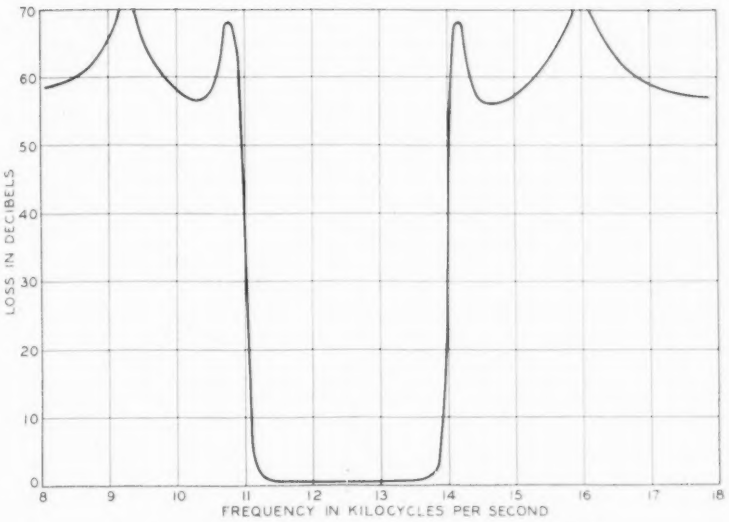


Fig. 17—Attenuation characteristic of a band pass filter.

of modifying the theoretical spacings for the transition factors has been to introduce peaks of loss near the band edges. The shortened intervals adjoining the cut-offs produce tanh $\theta/2$ curves which rise rapidly beyond these points to maxima slightly greater than unity, instead of approaching unity monotonically.

Having thus located the critical frequencies, we may readily complete the design of the filter in lattice form.

The formulation of the transfer-constant expression results in

$$\tanh \frac{\theta}{2} = K_1 \frac{\sqrt{1 - \frac{f^2}{f_a^2}} \left(1 - \frac{f^2}{f_2^2}\right) \cdots \left(1 - \frac{f^2}{f_{10}^2}\right) \sqrt{1 - \frac{f^2}{f_b^2}}}{\left(1 - \frac{f^2}{f_1^2}\right) \left(1 - \frac{f^2}{f_3^2}\right) \cdots \left(1 - \frac{f^2}{f_9^2}\right) \left(1 - \frac{f^2}{f_{11}^2}\right)},$$

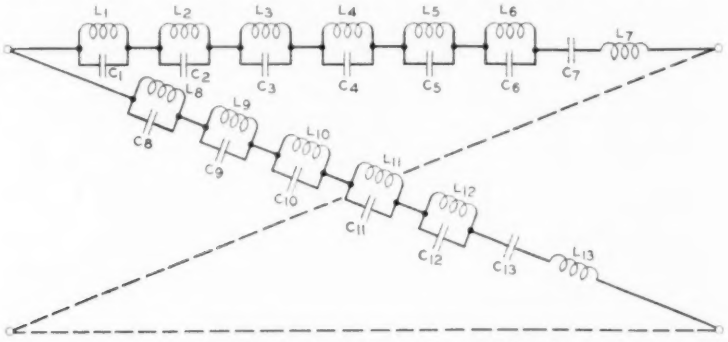
where f_a and f_b represent the cut-offs, and the other f's intervening critical frequencies in order of magnitude, as shown by Table IV.

A suitable form for the image impedance must next be obtained and since it is normally determined by requirements with which we are not now concerned, we will adopt the simplest possible expression, namely

$$Z_I = K_2 \frac{\sqrt{1 - f^2/f_a^2} \sqrt{1 - f^2/f_b^2}}{if},$$

where K_2 is determined from the condition that $Z_I = 600$ ohms when

$f = \sqrt{f_a f_b}$. The impedance functions Z_x and Z_y are now readily found by means of (1) and (2), and with the help of Foster's formula the element values can be obtained. These are shown in Fig. 18.



$L_1 = 0.0675$ mh.	$C_1 = 3.035$ mf.
$L_2 = 0.6529$ mh.	$C_2 = 0.2810$ mf.
$L_3 = 0.1958$ mh.	$C_3 = 0.8622$ mf.
$L_4 = 0.1770$ mh.	$C_4 = 0.8801$ mf.
$L_5 = 0.1330$ mh.	$C_5 = 1.085$ mf.
$L_6 = 0.0404$ mh.	$C_6 = 3.325$ mf.
$L_7 = 35.66$ mh.	$C_7 = 0.0046$ mf.
$L_8 = 0.1620$ mh.	$C_8 = 1.182$ mf.
$L_9 = 0.2021$ mh.	$C_9 = 0.8703$ mf.
$L_{10} = 1.335$ mh.	$C_{10} = 0.1214$ mf.
$L_{11} = 0.1097$ mh.	$C_{11} = 0.9112$ mf.
$L_{12} = 0.1089$ mh.	$C_{12} = 1.276$ mf.
$L_{13} = 35.86$ mh.	$C_{13} = 0.0046$ mf.

Fig. 18—Band pass filter.

This example illustrates the way in which the analysis may be applied to a typical problem in network design. The practical design would not ordinarily be complete at this point, however, since, as was mentioned previously, it is seldom desirable actually to construct the network as a single symmetrical lattice. Improved stability with respect to variations of the elements from their design values is obtained if the lattice is resolved into its components, that is, the elementary lattice sections which when operated in tandem have the same transmission properties. This question is discussed in a recent paper.[3] Furthermore, unbalanced structures equivalent to the symmetrical lattice but employing fewer elements are known,[23] and expense can usually be reduced by resorting to one of these.

[3] H. W. Bode, loc. cit. It may be interesting to observe that in the terminology of that paper the elementary constituents of linear phase shift filters are usually complex m sections.

[23] A linear phase shift lattice filter cannot, of course, be constructed as a sequence of Π or T sections, but equivalences in generalized bridged-T configurations exist. General equivalences in configurations employing ideal transformers are familiar in the literature. See, for example, Cauer, loc. cit., or Jaumann, *E. N. T.*, July, 1932, p. 243.

Part III—Filters with Linear Phase Shift Through the Cut-off

It was the conclusion of the theoretical discussion that any desired approximation to ideal filter characteristics may be obtained from a finite network, so long as a finite transition interval separates transmitting from attenuating bands. The transition interval can be taken small at pleasure, but very small transition intervals are associated with networks of many natural frequencies and numerous elements. We have already seen how considerable economies in meeting a given attenuation requirement could be obtained if the phase requirement were subordinated or removed entirely. We now consider the contrary case, in which major emphasis is placed upon the phase characteristic of the filter. Filters of this type are of practical interest in picture transmission systems since instruments used in the reproduction of images seem to be much more sensitive to the effects of phase distortion than the ear. The selectivity required from filters used in such systems is comparatively modest, but phase linearity is required not only in the practical transmission band but also through the transition interval into the region of rising attenuation.

In one important particular the present problem differs from those previously considered. In the present analysis we can no longer regard the adjustment of Z_1 and the adjustment of θ as independent problems. On the contrary, in the attenuating region the contribution of θ to the phase shift is constant and we must therefore rely upon reflection effects to maintain the desired linear characteristic. Moreover, near the cut-off θ must be very carefully adjusted with respect to Z_1 in order that the contributions to the phase shift from reflection and interaction effects may preserve the linearity through the transition band also. The added restrictions imposed by the extension of the phase requirement require a revision of the frequency spacings already found, and set limits upon the approximation to ideal characteristics obtainable from reactive networks of reasonable complexity.

Use of Reflection Effects to Produce Linear Phase

In the practical transmission band, Z_1 can be adjusted to approximate R sufficiently closely to make reflection and interaction effects negligible. Therefore, in this range the total insertion phase is the same as the transfer constant phase, and, as before, is to be obtained from a chain of uniformly spaced critical frequencies in $\tanh \theta/2$. In the practical attenuating band, on the other hand, we find that the imaginary part of θ is either 0 or π, while interaction effects can be

ignored if we assume the loss to be reasonably high. The variation of the phase shift with frequency must therefore be attributed to reflection effects, which we can write as

$$e^{\theta_r} = \frac{\left(1 + \frac{Z_I}{R}\right)^2}{4\frac{Z_I}{R}},$$

where θ_r is the sum of the reflection effects at the two ends of the structure.

Since Z_I is reactive in the attenuating band, the angle of the denominator in this equation is $\pm \pi/2$, while that of the numerator is 2 arctan Z_I/iR. Thus

$$B_r = \mp \frac{\pi}{2} + 2 \arctan \frac{Z_I}{iR}. \tag{25}$$

Except for the constant term, which we will consider presently, this is a function of precisely the type we have been considering. Hence if the impedance controlling factors are spaced at the same uniform interval that was used in the pass-band, the phase slope will be constant and equal in both bands.

Phase Characteristics in Transition Intervals

The transition factors—or rather, factor, since clearly we have to rule out the solutions for $m > 1$—must be determined so that these linear parts of the phase characteristic are joined by a chord of the same slope. If we suppose the transition interval to be bounded by the last uniformly spaced frequencies of the transfer constant and image impedance chains, and to contain only the cut-off factor, it is easily shown that it must include a net change in phase of $3\pi/2$ radians. The interval must therefore contain 3/2 uniform spaces if the average slope is to be correct. Considerations of symmetry to be described later require that the cut-off be the center of the interval, which thus comprises two three-quarter spaces. The behavior of the several components of the total insertion phase is exhibited by Fig. 19, in which B, B_r, and B_i refer respectively to the phase shifts contributed by the transfer constant, by reflection effects, and by the interaction factor. The mutually annulling discontinuities of $\pi/2$ radians in B_i and B_r at the cut-off are noteworthy.

The fact that this choice of parameters is sufficient as well as necessary to obtain the desired linearity of phase shift is not easily shown analytically. It can, however, be verified by direct computa-

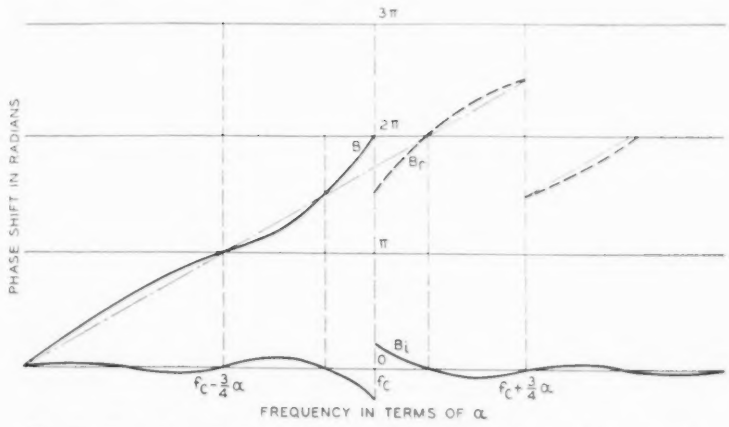


Fig. 19—Transfer, reflection, and interaction phase in the transition interval.

tion. For this purpose the customary resolution of the total insertion loss into transfer constant, reflections, and interaction is not very useful because of the indeterminacies found at the cut-off. This difficulty is avoided by expressing Z_I and θ in terms of the lattice impedances, in which event

$$e^{\gamma} = \frac{1 + \frac{Z_x Z_y}{R^2} + \frac{Z_x}{R} + \frac{Z_y}{R}}{\frac{Z_x}{R} - \frac{Z_y}{R}}, \tag{26}$$

where γ is the total insertion loss.

If iX_x and iX_y be written for Z_x and Z_y, the insertion loss and phase shift are given by

$$\tan B_\gamma = \frac{X_x X_y - R^2}{R(X_x + X_y)} \tag{27}$$

and

$$e^{A\gamma} = \frac{\sqrt{(R^2 + X_x^2)(R^2 + X_y^2)}}{R(X_x - X_y)}. \tag{28}$$

Equation (27) can be used to confirm our previous choice of the location of the cut-off. At this frequency one of the two reactances, X_x and X_y, will be either resonant or anti-resonant. It is evident from (27) that if the phase shift is to have the desired value, $(n + 3/4)\pi$, at the assumed cut-off the non-resonant impedance must have the magnitude R. That this value is approximated follows from the symmetrical spacing of transfer constant and image impedance

controlling frequencies with respect to the cut-off. On both sides of the transition interval, in the regions of uniform spacing of poles and zeros, the non-resonant reactance approximates $R \tan \pi f/2\alpha$ or $R \cot \pi f/2\alpha$ and at the middle of each space, where $\pi f/2\alpha$ is an odd multiple of $\pi/4$, is numerically equal to R. Hence, by symmetry, this must also be the value approximated at the middle of the non-uniform interval between the two chains, i.e., at the cut-off frequency.

Nature of the Approximation

The argument of Part I shows that the three-quarter spacing between the cut-off and the chain of transfer constant controlling factors results in poorer approximations to phase linearity in the transmission band and to complete suppression in the attenuating band than would the half-spaced cut-off solution. The three-quarter spacing between the cut-off and the chain of impedance controlling frequencies also leads to less perfect uniformity of the impedance characteristic. This is the price we pay for the larger range of phase linearity. Nevertheless, the error of approximation for both θ and Z_I if we follow the sense of equation (22) can be shown to be $\frac{1}{8}\alpha\left(\frac{1}{1-f}-\frac{1}{1+f}\right)$, when α is small, and hence can be made as small as we please by a suitable choice of α.[24] So far as the phase and impedance characteristics are concerned, experience shows that satisfactory precision can be obtained with a moderate value of α. The situation with respect to the attenuation characteristic is more serious. As we have already seen, the best approximation in the attenuating band is obtained by a cut-off spacing which is, if anything, slightly less than, rather than slightly greater than, $\alpha/2$. Furthermore, it appears from the above formula that with the three-quarter cut-off spacing, the approximation error at a given frequency in the attenuating band is proportional only to the first power of α. Hence cutting α in two, which substantially doubles the number of elements in the structure, adds but 6 db to the attenuation at this frequency. It is clear that a practical limit is thus set upon the suppression which can be provided.

Since the attenuation of the structure is relatively low, the contribution of reflection effects to the total loss is correspondingly important. A peak of loss occurs at each impedance controlling frequency, where the lattice impedances are zero or infinite together.

[24] It is not true that the error in $\partial\beta/\partial\omega$ vanishes with α. However, in the following example, which may be taken as typical, the variation of $\partial\beta/\partial\omega$ is still only about 1 per cent of its average value.

At these frequencies the image impedance changes sign, and therefore also the constant term of equation (25). Thus, although the phase slope is uniform throughout the attenuating range, the phase characteristic itself suffers discontinuities of π radians at each impedance controlling frequency. Whether the discontinuity is an increase or a decrease of π radians is not distinguishable for a non-dissipative network. When parasitic dissipation is taken into account the peaks are finite and the phase increases or decreases according as the line- or cross-arm of the lattice has the smaller resistance component at the peak frequency. The infinite peak at this frequency, and the associated abrupt change in phase, can evidently be restored by adding additional resistance to the smaller impedance so as to bring the arms into balance.

This observation is of importance in considering the effect of dissipation on the phase shift. A counterpart of Mayer's theorem can be found which relates the change in phase shift resulting from uniform dissipation in the network elements to the slope of the loss curve. The formula is

$$\Delta B \doteq - \omega d \frac{\partial A}{\partial \omega},$$

where d is the dissipation constant, and where A and B are in nepers and radians respectively. In the transition interval, where the slope of the loss curve is great, the effect of uniform parasitic dissipation may reduce the phase appreciably. This effect can be compensated by small modifications in the theoretical frequency spacings, or by the introduction of a lumped resistance to balance the bridge at the first impedance controlling frequency, according to the plan suggested above.

Example

To illustrate the performance of this sort of network, we may consider a low-pass filter containing four evenly spaced critical frequencies in the practical transmission band. Subsequent natural frequencies will then occur at 4.75α, 5.5α, 6.5α, etc., according to the rule for three-quarter spacing adjacent to the cut-off. We may suppose that the requirement for linearity of phase shift does not extend above 7.5α, so that the sequence of uniformly spaced impedance controlling frequencies may be terminated after this point according to the scheme proposed in the case of the high-pass filter. In the frequency range of interest, we can replace the omitted chain of uniformly spaced frequencies by a single natural frequency at double spacing. The transfer constant and image impedance expressions can then be written as:

$$\tanh \frac{\theta}{2} = i \frac{\pi f}{2\alpha} \frac{\left(1 - \frac{f^2}{(2\alpha)^2}\right)\left(1 - \frac{f^2}{(4\alpha)^2}\right)}{\left(1 - \frac{f^2}{\alpha^2}\right)\left(1 - \frac{f^2}{(3\alpha)^2}\right)\sqrt{1 - \frac{f^2}{(4.75\alpha)^2}}},$$

and

$$Z_I = R \frac{\sqrt{1 - \frac{f^2}{(4.75\alpha)^2}}\left(1 - \frac{f^2}{(6.5\alpha)^2}\right)\left(1 - \frac{f^2}{(9.5\alpha)^2}\right)}{\left(1 - \frac{f^2}{(5.5\alpha)^2}\right)\left(1 - \frac{f^2}{(7.5\alpha)^2}\right)}.$$

These equations determine Z_x and Z_y, the values of which may be used in equations (27) and (28) to calculate the performance. This is shown by Fig. 20, after dissipative effects have been taken into account.

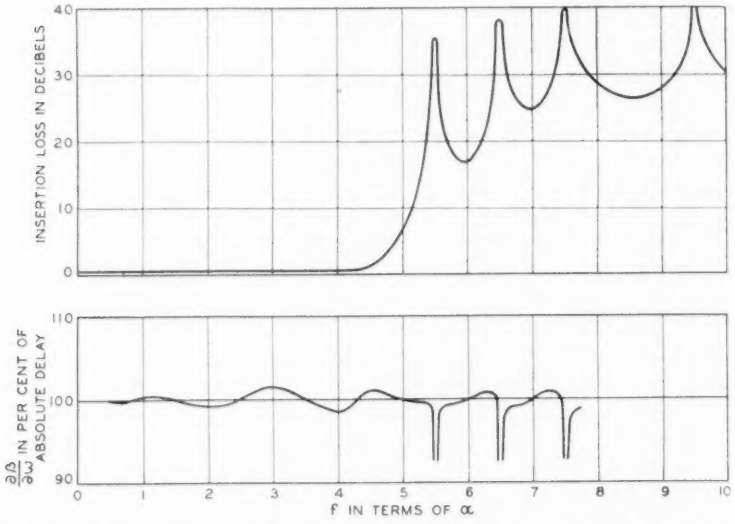


Fig. 20—Performance of a low pass filter having linear phase through the cut-off.

The approximation of the phase characteristic to linearity has again been indicated by exhibiting departures of the slope from the average.

It is observed that the approximation obtained by the three-quarter spacing is as close in the transition interval from 4α to 5.5α as in the practical transmitting band below 4α. In practical design problems, the phase shift is unlikely to be of interest beyond the first or second reflection peak, so that the chain of impedance controlling frequencies might be sooner terminated.

The loss characteristic reveals that no very high degree of suppression is attained. In fact, the loss falls to about 16 db in the trough beyond the first reflection peak. So serious a prejudice in favor of the phase characteristic would render the design unsuitable for certain engineering purposes. There are open, however, several possibilities for increasing the attenuation. Small modifications in the theoretical design parameters of the type which have been described, and in particular, slight separations of the theoretically coincident impedance controlling frequencies in the two arms of the lattice, enable the loss to be somewhat improved without much degradation of the phase characteristic. If very much higher attenuation is demanded, it can be provided by two simple structures of this type, separated by a resistance pad to preserve the reflection effects upon which the phase characteristic depends.

Further possibilities are suggested by combination of two principles already developed. It has been observed that a reduction in the three-quarter spacing of the cut-off would improve the selectivity of the structure but would also unduly increase the slope of the phase characteristic in the transition interval. We have also seen, however, that the result of uniform dissipation in the network elements is to diminish the phase shift in this region. Hence our analysis suggests that we may be able to obtain the desired phase characteristic in conjunction with the shorter cut-off spacing necessary for high selectivity if we deliberately increase the dissipation in the network.

A concomitant result of such procedure is seen to be an increase in the uniform loss in the transmission band, which may not always be desirable. Neither does the attempt to provide the phase property without sacrifice of high loss through the introduction of uniform dissipation represent the most effective attack on the problem. To achieve this end, resistances must be associated with the reactive elements of the lattice impedances in a precisely determined manner, not to be deduced solely from the foregoing theory of reactive networks. The elaboration of the theory to include also resistive impedance elements serves to determine a filter whose attenuation changes continuously from a low, uniform value in the pass-band to an arbitrary value in the attenuation bands with linearity of phase shift and, in addition, the third ideal property of constant impedance. The general theory, however, can more appropriately form the subject of a subsequent paper.

The solution of this problem completes the application of the methods for realizing ideal filter properties. We have seen that if all

the properties are of importance and the desired approximations close, we are led to networks which, while formally simple, involve correspondingly numerous natural frequencies. On the other hand, if the impedance, or the phase, or the loss property be subordinated in respect to the others, suitable modification of the analysis allows the remaining properties to be realized with simplification of the structure.

Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics *

By C. R. BURROWS, L. E. HUNT and A. DECINO

This paper, a sequel to one entitled "Ultra-Short-Wave Propagation," deals with transmission in urban areas. The experimental data were obtained in the City of Boston on a frequency of 34.6 megacycles per second by means of a specially equipped motor car for carrying the receiver while the transmitter was fixed or vice versa.

Mass plots of these data show that the mean field strength varies inversely as the square of the path length which is the same variation as would be expected for level terrain in the absence of buildings. The same data are presented in the form of field strength contour maps.

These data are interpreted on the basis of the same physical picture which has been established for open country. The present data preclude interpretation upon the basis employed by earlier investigators of ultra-short-wave propagation through urban areas.

It is concluded that ultra-short-wave transmission in urban territory may be interpreted on the basis of transmission over level land plus the wave interference patterns caused by reflections from the buildings and an additional attenuation which on the average is independent of the length of the transmission path. Also, if the theoretical formula for the propagation of ultra-short waves over level terrain is used to calculate the received field in urban territory, and the height of the fixed antenna is measured from the local roof level instead of from the ground, these data indicate that the field strength so calculated would be near to the mean of the actual received field strengths in urban territory.

Introduction

THIS paper is a sequel to an earlier paper, "Ultra-Short-Wave Propagation," [1] which dealt mainly with transmission across open country. In the present paper the research has been extended to include transmission within a built-up region. Additional problems of transmission within urban areas that result from man-made interferences, such as the noise produced by automobile ignition systems, have been investigated.

A specially equipped motor car was used as a mobile laboratory for most of this work both because of its convenience as a means of obtaining transmission data and because of the importance of mobile communication itself.

This paper describes general characteristics and quantitative measurements of the received signal on 34.6 megacycles. Transmission

* Published in *Electrical Engineering*, January, 1935.

[1] J. C. Schelleng, C. R. Burrows and E. B. Ferrell, "Ultra-Short-Wave Propagation, *Proc. I. R. E.*, Vol. 21, pp. 427–463, March, 1933 and *Bell Sys. Tech. Jour.*, Vol. 12, pp. 125–161, April, 1933.

phenomena were studied in both directions between a fixed location on a building and the mobile laboratory.

Apparatus

Both terminals employed vertical half-wave antennas which were connected to balanced circuits by means of symmetrical two-wire transmission lines. At the fixed locations unloaded antennas were used; at the mobile terminal, in order to limit their heights to eight feet above the ground and maintain the symmetry the antennas were loaded so that their lengths were reduced to about a quarter of a wave-length (Fig. 1).

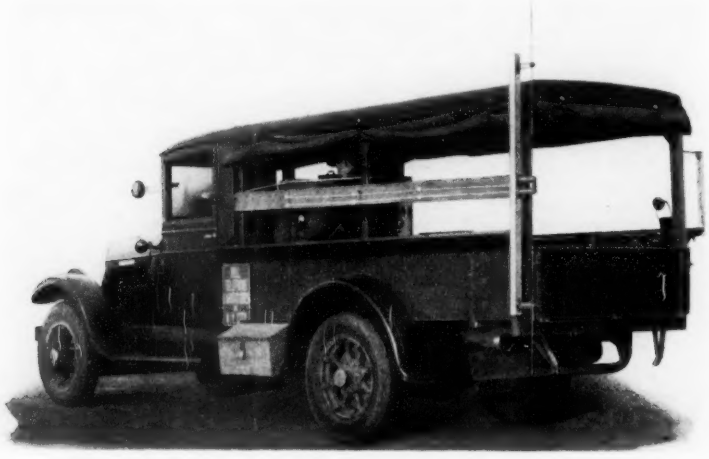


Fig. 1—Mobile receiving equipment.

The transmitter consisted of an electric oscillator employing two 75-watt tubes operating in push-pull relationship. At the fixed location, where ample power was available, the transmitter was capable of producing one ampere [2] of 100 per cent modulated carrier in its antenna without undue distortion. The mobile transmitter which used a dynamotor for tube plate supply was capable of producing the same current in its antenna. This corresponds to about six decibels less power due to the shorter antenna length.

The measuring set was of the double detection type with balanced high-frequency circuits, push-pull first detector and calibrated inter-

[2] The current was measured by a Weston type 425 thermoammeter at the current maximum.

mediate frequency attenuator.[3] This receiving equipment was calibrated in absolute units by a method described in the appendix. A mechanism for recording the field strength was attached to the measuring set: this consisted of a roll of paper that could be driven either by clock-work or by the rear wheels of the truck. The position of the recording pen was controlled by the setting of a manually operated variable attenuator. Samples of the type of record obtained are shown in Figs. 6 and 7.

Locations

The radiator for the fixed transmitter was supported by a fifty-foot pole above the roof of a seven-story building at the corner of Berkeley and Stuart Streets in the business section of Boston. The building is about 90 feet high, making the center of the antenna about 130 feet above the ground. Thus, the antenna was higher than most of the buildings of the city though it was lower than a few buildings nearby.

The antenna for the fixed receiver was supported by a 20-foot pole from the middle of the highest ridge of a gabled building making the center of the antenna about 80 feet above the street level. This building is located on the side of a slight slope in a fairly heavily wooded territory, on Seaverns Street near Center Street.

Field Strength Measurements

Transmitter at a Fixed Location

With a current of one ampere in the half-wave antenna[4] above the building at Berkeley and Stuart Streets and with the receiver in the truck, field strength measurements were made along various routes throughout Boston. These data have been averaged by one-tenth-mile intervals when the average radial distance was less than two miles and by half-mile intervals for greater average distances. A plot of these data is shown in Fig. 2. The points lie approximately on an inverse-square-of-distance line with deviations ranging up to about $\pm$ 10 db. An effort has been made to separate the points taken in the high building area. These points (shown as open circles) lie somewhat below the others with a few particularly low field strengths. The lowest field strengths of the business district were measured along the shore near Charles River Dam and near State Street on Atlantic Avenue. The field strengths in the business district would be expected to be lower because of the presence of the high buildings. The lower

[3] The set was similar to that described by Friis and Bruce, *Proc. I. R. E.*, Vol. 14, pp. 507–519, August, 1926.

[4] Since the antenna was in free space in so far as radiation resistance is concerned, this corresponds to a radiated power of 73 watts.

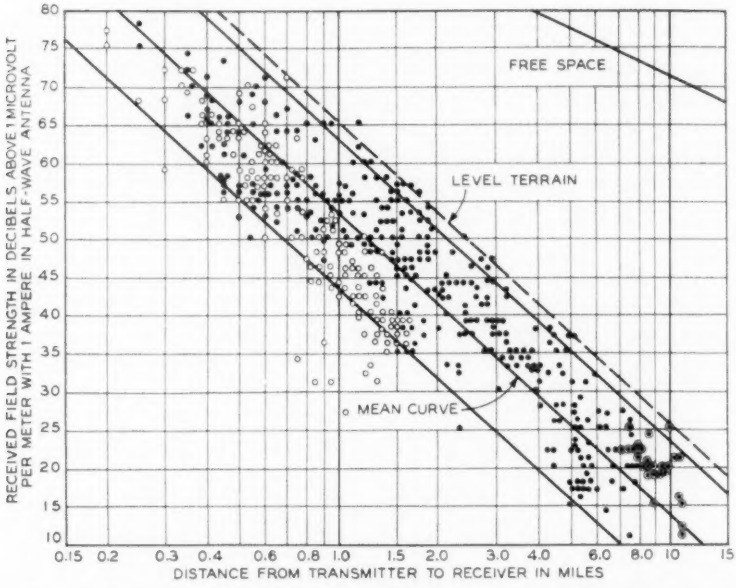


Fig. 2—Mass plot of field intensities measured at various distances from the transmitter at Berkeley and Stuart Streets in Boston. The values corresponding to distance less than two miles represent field strengths averaged over one-tenth mile intervals, while those for greater distances represent averages over one-half mile intervals. The open circles indicate fields in the high building area. Residential points outside the city limits have been enclosed in circles.

residential field strengths correspond to the region beyond Chestnut Hill.

When attempting to interpret the results of the mass plot of Fig. 2 a natural method would be to assume transmission as in free space plus an additional attenuation due to the proximity of the earth and obstacles above the earth's surface. The simpler case of transmission over level terrain in the absence of obstacles will be considered first.

It has been experimentally determined that the propagation of ultra-short waves over unobstructed paths follows the laws of optics [1, 5] so that the resultant field is composed of a well-defined reflected wave superposed upon a direct wave. Consequently for propagation over level terrain, the explanation is as follows (Fig. 3): Energy is propagated from a transmitter at A, at a height of h_1 above the ground, to a receiver at B, at a height of h_2 above the ground, both directly, as

[1] Loc. cit.

[5] C. B. Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proc. I. R. E.*, Vol. 21, pp. 764–801, June, 1933.

represented by r_1, and by reflection at G, as represented by r_2, the distance between transmitter and receiver being represented by d. For the practical case where h_1 and h_2 are small compared with d the reflected wave impinges upon the ground at nearly grazing incidence, so that a negative reflection coefficient the magnitude of which is unity [6] for ordinary ground (not water) is obtained. This

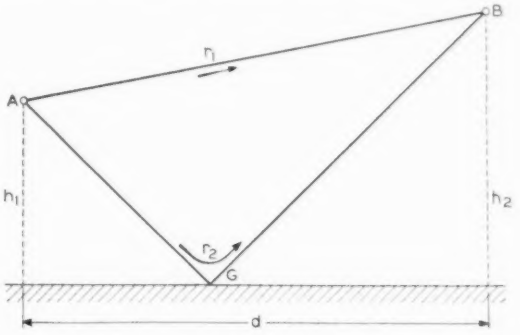


Fig. 3.

results in the field at B being the difference between two vectors of approximately equal magnitude and differing in phase by an amount corresponding to the difference in path lengths, r_2 and r_1. For the case under consideration,

$$r_2 - r_1 = 2h_1h_2/d, \tag{1}$$

and the angle between the vectors is

$$2\pi(r_2 - r_1)/\lambda = 4\pi h_1h_2/\lambda d. \tag{2}$$

[6] The magnitude is more exactly $1 - 2\epsilon(h_1 + h_2)/d\sqrt{\epsilon - 1}$ for vertical polarization and $1 - 2(h_1 + h_2)/d\sqrt{\epsilon - 1}$ for horizontal polarization, making the corresponding values for the received fields,

$$E_0\left(\frac{4\pi h_1h_2}{\lambda d}\right)\sqrt{1 + \frac{\epsilon^2(h_1 + h_2)^2\lambda^2}{(\epsilon - 1)4\pi^2h_1^2h_2^2}} \tag{a}$$

and

$$E_0\left(\frac{4\pi h_1h_2}{\lambda d}\right)\sqrt{1 + \frac{(h_1 + h_2)^2\lambda^2}{(\epsilon - 1)4\pi^2h_1^2h_2^2}} \tag{b}$$

respectively, instead of as in equation (3). When the lower of the two antennas is more than a couple of wave-lengths off the ground, the radicals are substantially unity. For the case under consideration it would be more accurate to refer to expression (*a*) as the theoretical formula but lack of knowledge of the magnitude of the dielectric constant and antenna heights that apply would introduce unnecessary uncertainty if the results were referred to this formula. It might be remarked that neglecting the presence of buildings and referring all heights to the local street level the radical represents an increase of 7–12 db for vertical polarization in the cases under consideration.

Except in the immediate proximity of the transmitter this angle is small and the resultant field is

$$E = E_0(4\pi h_1 h_2/\lambda d), \tag{3}$$

where E_0 is the free space field. Since [7]

$$E_0 = \frac{60\pi II}{\lambda d}, \tag{4}$$

the resultant field becomes

$$E = 240\pi^2 II h_1 h_2/\lambda^2 d^2. \tag{5}$$

Equation (5) shows that the field over level terrain is inversely proportional to the square of the distance from the source.[8]

Data presented in Fig. 8 of reference 1 show that at a frequency of 69 megacycles per second the field strength variation with distance follows approximately the inverse-square relationship for a range of from two to ninety kilometers.[9] In fact, the best straight line through these data agrees with the numerical values obtained from equation (5) well within the accuracy of the experimental data. Experiments designed to test the validity of this equation now are being conducted at Deal, New Jersey and data obtained to date confirm it both as to absolute value and variation with terminal heights and wave-length for horizontal polarization within the range $2 < h_1 < 25$, $2 < h_2 < 25$, $2 < \lambda < 17$, $d = 9{,}420$ and $26{,}300$, all measured in meters. The experimental confirmation of this formula for these distances indicates that the effect of the earth's curvature is secondary to the negative reflection effect upon which this formula is based. This might be expected in view of the fact that both diffraction and refraction tend to mitigate the additional attenuation that would be caused by reflection from a plane tangent to the earth's surface at the point of geometric reflection.

[7] If I is in amperes, d in meters, and H the effective height of the antenna, and λ the wave-length in the same units, E_0 is given in volts per meter.

[8] Since distance appears in this equation only as a factor and not as an exponent, the reduction with distance of the field strength of ultra-short waves over level terrain is independent of wave-length, polarization, dielectric constant, etc., as all of these quantities cancel in the ratio of the field strength at one point to that at another. The absolute magnitude of the field strength is proportional to the frequency for the same radiated power and antenna heights. If the antenna heights are sufficiently low, the field is also dependent upon the polarization and ground constants as indicated by expressions (*a*) and (*b*) of footnote 6.

[9] While undoubtedly at the greater distances the field suffers additional attenuation above that shown by equation (5) due to the curvature of the earth, such additional attenuation evidently takes place at distances beyond those employed in any of the authors' experiments.

When the propagation is through built-up areas instead of over level terrain the condition is more complicated. Even here, however, for terminals well above the tops of buildings, theoretical considerations [1] indicate that the same explanation of direct and reflected waves is valid. Data presented by Jones [10] may be used as a verification for this explanation even for transmission over buildings. Fig. 11 of his paper shows that for heights between 170 and 1,500 feet the field is proportional to the height in accordance with equation (5). When the terminals are lowered within the building region, the field should decrease more rapidly than proportionally with the height above the ground. In fact, data presented in Fig. 4 indicate that with one

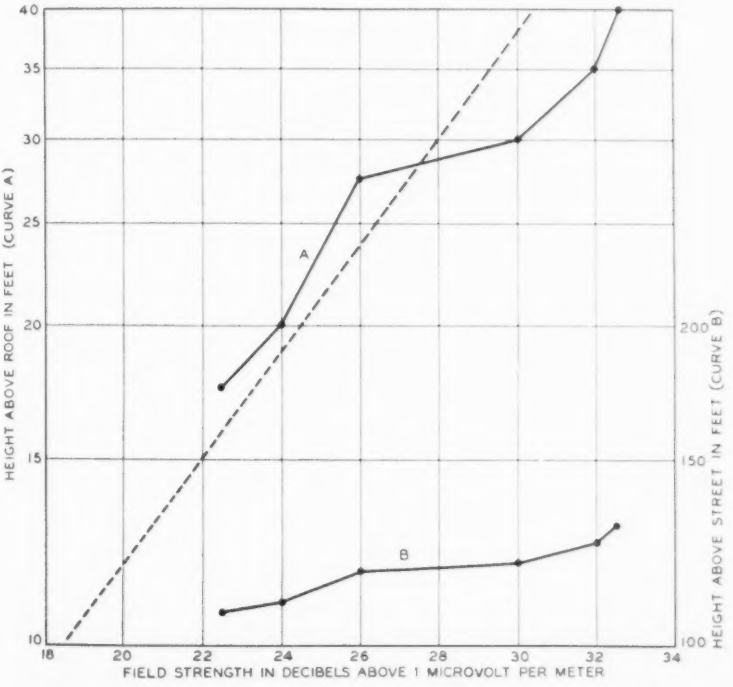


Fig. 4—Variation of field received at Berkeley and Stuart Streets with antenna height. Curve *A* shows the variation with height above the roof while Curve *B* shows the variation with height above the ground. The slope of the broken line indicates a linear relationship between field and height.

[1] Loc. cit.

[10] L. F. Jones, "A Study of the Propagation of Wave-lengths between Three and Eight Meters," *Proc. I. R. E.*, Vol. 21, pp. 439–486, March, 1933.

terminal above a flat roof that was approximately the same height as other nearby flat roofed buildings, the field strength is more nearly proportional to the height above the roof than to the height above the ground. When the terminals are lowered below the average building height additional complications are introduced. While this is somewhat difficult to picture because of the irregularity of the surface bounding the transmitting medium [11] the main outline seems simple enough. Above the building level there is a tendency for the field to follow the simple rules that hold for transmission over level country. In general the field strength actually received in the street would be proportional to the field strength overhead but of smaller amplitude since it is a product of scattering. This does not imply that the street signal comes down vertically; it probably is the result of scattering from points lying in a fairly large zone about the receiver and consists of a multiplicity of signals traveling in inclined directions.

Returning now to the present data on the propagation of ultra-short waves through urban areas, Fig. 2 shows [12] that the field strength is in general inversely proportional to the square of the distance from the transmitter. The mean curve through the data is 12 db below the curve for level terrain free from obstacles, plotted from equation (5) above, indicating the additional attenuation due to man-made structures. An analysis of the individual points shows that the reduction in field due to the obstacles (i.e., in addition to the level terrain attenuation) is independent of the distance so that there is no *absorption* due to the buildings in the usual meaning of the word; otherwise the additional attenuation would increase with the distance.

This method of interpretation is radically different from that of investigators of the propagation of ultra-short waves through urban areas whose papers have come to the attention of the authors.[13, 14, 15] They have assumed that the transmission occurs as in free space except for an additional attenuation *through the absorbing layer of buildings*. Such an assumption that the propagation of ultra-short waves is unaffected by the presence of the ground except in so far as the waves penetrate the absorbing layer of buildings, appears to be

[11] This surface is, of course, that formed by the ground and the walls and tops of buildings.

[12] It will be shown later that the empirical formula assumed by Schröter, Sohnemann, Jones, and Muyskens and Kraus cannot be made to fit these data.

[13] F. Schröter, "Zur Frage des Ultrakurzwellen-Runkfunks," *E. N. T.*, Vol. 8, pp. 431–436, October, 1931.

[14] K. Sohnemann, "Feldstarkemessungen im Ultrakurzwellengebiet," *E. N. T.*, Vol. 8, pp. 462–467, October, 1931.

[15] Henry Muyskens and John D. Kraus, "Some Characteristics of Ultra-High-Frequency Transmission," *Proc. I. R. E.*, Vol. 21, pp. 1302–1316, September, 1933.

inconsistent with the physical picture [1, 5, 16, 17] of ultra-short wave propagation which has been confirmed by basic experimental data.

Since the inverse-square-of-distance relationship (equation 5) which results from this physical picture is so different from the exponential relationship which results from the *absorption* assumption of previous investigators, the question arises as to the possibility of reinterpreting their data on the basis of this physical picture. The data presented by Muyskens and Kraus as Fig. 2 of reference 15 has been replotted in Fig. 5 of the present paper on logarithmic coordinates in order to

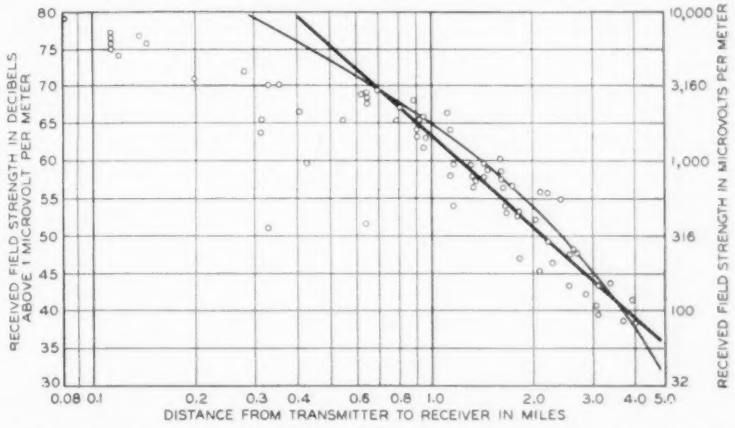


Fig. 5—Attenuation curve for 5 meter transmission as replotted from paper by Muyskens and Kraus, *Proc. I. R. E.*, Vol. 21, p. 1306, Sept., 1933. The straight heavy line shows an inverse-square of distance variation in accordance with the physical picture. The thin curved line is a replot of the curve presented by Muyskens and Kraus as representing these data by a variation according to an inverse-distance times an exponential factor. It is evident that these data may be interpreted equally well on the basis of the physical picture (heavy curve) as on the basis of the empirical equation assumed by Muyskens and Kraus.

facilitate reinterpretation on the basis of this physical picture. Figure 5 shows that it is possible to interpret their data as following an inverse-square-of-distance law equally as well as an inverse-distance law times an exponential factor.[18] In this interpretation little weight

[16] Bertram Trevor and P. S. Carter, "Notes on Propagation of Waves Below Ten Meters in Length," *Proc. I. R. E.*, Vol. 21, pp. 387–426, Mar., 1933.

[17] Carl R. Englund, Arthur B. Crawford and William W. Mumford, "Some Results of a Study of Ultra-Short-Wave Transmission Phenomena," *Proc. I. R. E.*, Vol. 21, pp. 464–492, March, 1933 and *Bell Sys. Tech. Jour.*, Vol. 12, pp. 197–227, April, 1933.

[18] Comparison of Figs. 2 and 5 on an absolute basis is difficult. The lower density of the buildings in Ann Arbor and the fact that the fixed antenna was located in as open a space as possible at the corner of the roof, combine to reduce the effect of the buildings in lowering the mean field strengths. Also the apparent absence of a measurement of the radiated power in the Ann Arbor experiments precludes the determination of the absolute value of the attenuation from the experimental data.

has been given to the points which are well below the curve, in accordance with the view of the experimenters that these points represent particularly unfavorable receiving locations.

Trevor and Carter [16] have made a similar interpretation of the data presented by Jones [10] which shows that for the larger distances the field strength was inversely proportional to the square of the distance. If this inverse-square-of-distance curve were extended to shorter distances it would be found that most of the nearby points would lie somewhat below it. This is presumably because of the lack of favorable receiving locations in the high-building area. While the empirical formula arrived at by Jones may represent his data satisfactorily, the physical picture assumed of a free space field times an absorption factor is untenable since it requires a radiated power approximately 20 db below that measured. Undoubtedly, the power radiated is not in error by this amount, since Trevor and Carter obtained a satisfactory numerical check on the basis of the other picture by using the value of power radiated as given by Jones.

It is possible, of course, to represent any data by an inverse-distance factor times an exponential factor for a limited range of distances. An attempt to do this with the data of Fig. 2 by making the empirical curve agree with the experimental inverse-square-of-distance curve at 1 and 4 miles results in a curve that agrees well with the data between 0.6 and 5.0 miles but is 11 and 22 db low at 0.2 and 12.0 miles, respectively. Even if these rather large discrepancies at the limits of the curve were neglected it would still be impossible to interpret the data in terms of the free space field times an exponential absorption factor, because of the fact that the empirical curve so determined requires a radiated power 35 db below that measured; this is untenable since the over-all uncertainty in the absolute value of the measurements is only a few decibels.

It should be pointed out that each point of Fig. 2 represents the average field over an interval of either a tenth or a half mile depending upon whether the transmission path involved was less or greater than two miles. Within each interval the field varied by five to fifteen decibels because of the local wave interference pattern, as is shown by the samples of the graphs taken with the recorder which are presented in Figs. 6 and 7. Fig. 6 is an example of the record taken in the business district of Boston at a distance of about one and a half miles from the transmitter near the region *A* shown in Fig. 8. The maxima and minima are spaced very closely and differ by ten to fifteen decibels. This was characteristic of the type of record obtained at the shorter distances. At the greater distances the magnitude of the local variations was less, as illustrated by Fig. 7, which is a sample of the record

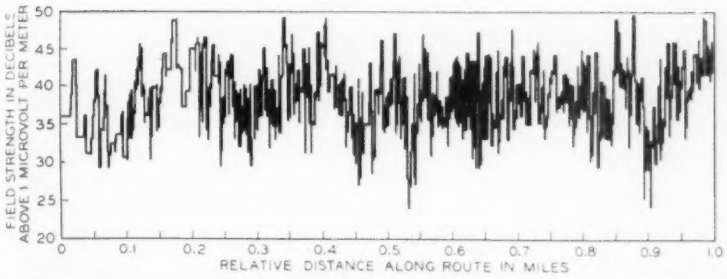


Fig. 6—Portion of record showing the large field strength variations as recorded while driving through the business district of Boston at a distance of about 1.5 miles from the transmitter.

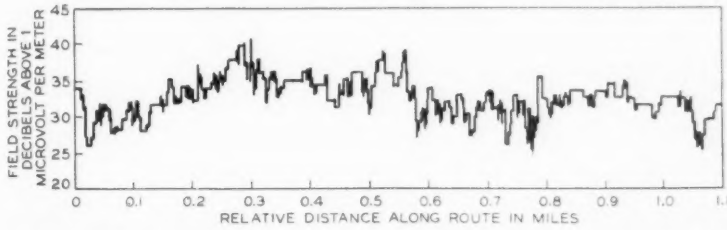


Fig. 7—Portion of record showing the small variations of field strength while driving through the residential section of Boston at a distance of about 5 miles from the transmitter.

taken at a distance of five miles (near *B* in Fig. 8). The change in the magnitude of the variations might have resulted from the fact that all of the data for the greater distances were taken in residential districts with correspondingly lower heights and densities of buildings.

An idea of the variations to be expected from the inverse-square-of-distance relationship is shown by the contour map of Fig. 8. The data already presented will give an idea of the impossibility of showing much detail in such a map. Likewise it would be an almost endless job to make measurements on every street in a city of this size. While data were obtained within reasonable intervals over the area for which solid contours are drawn (continuous field strength records taken over 143 miles of street are represented by this figure), another set of data might result in a somewhat different looking map. The broken contours are not based upon field strength measurements, but are merely a plausible way of joining the solid contours to aid the eye of the reader. The two 20 db contours in the lower part of the figure have not been joined because of insufficient data. The low field strengths along the Neponset River may be a result of local conditions and possibly the 20 db contours should be continued to the right along

an arc of more nearly constant radius. In this connection it may be mentioned that the field is nearly constant within the bulge of this contour to the southwest.

A striking fact brought out by the map is the crowding of the contours in the business district. There are particular directions for which the attenuation is greater presumably due to the combined effects of high buildings, for example to the east-northeast. There are other directions where the field strength is higher than the average. Several such places were noted when salt water extended immediately in front of the measurement location in the direction of the transmitter. The closed contours in the Mystic River to the north illustrate the better reception over salt water as predicted by theory.[1] An example of the records taken over bridges upon which these contours are based is shown in Fig. 9. (The route over which these data were taken is

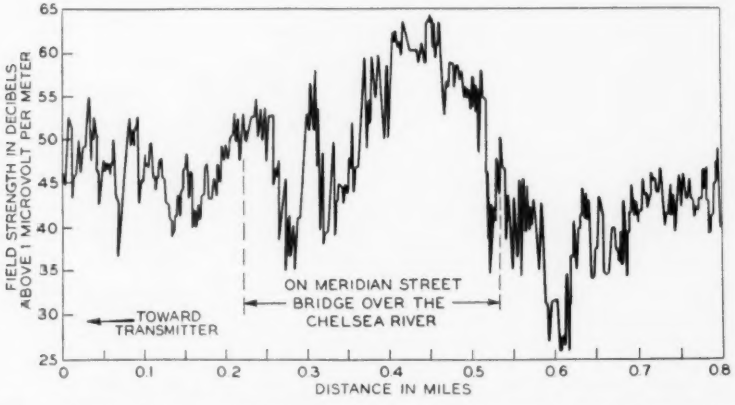


Fig. 9—Portion of records showing the variation of field strength on going over water.

indicated at *C* of Fig. 8.) The average increase in field when going over bridges was 10 db. This may be explained by the better conductivity of the water which results in a receiving directive characteristic that is more favorable to low angle reception. In some cases, the absence of buildings and increase in the height of the antenna due to the elevation of the bridge undoubtedly contributed to the greater field.

Another fact that is illustrated by the map is that the effects of obstacles and type of terrain are in general local. For example, the higher field strengths beyond salt water are soon reduced to normal with the intravention of additional land, illustrated at *D* in Fig. 8.

It was found that the field strengths along Columbus Avenue (indicated by *E* on Fig. 8) were about 10 db higher than those obtained on either side. This increase probably resulted from the fact that a

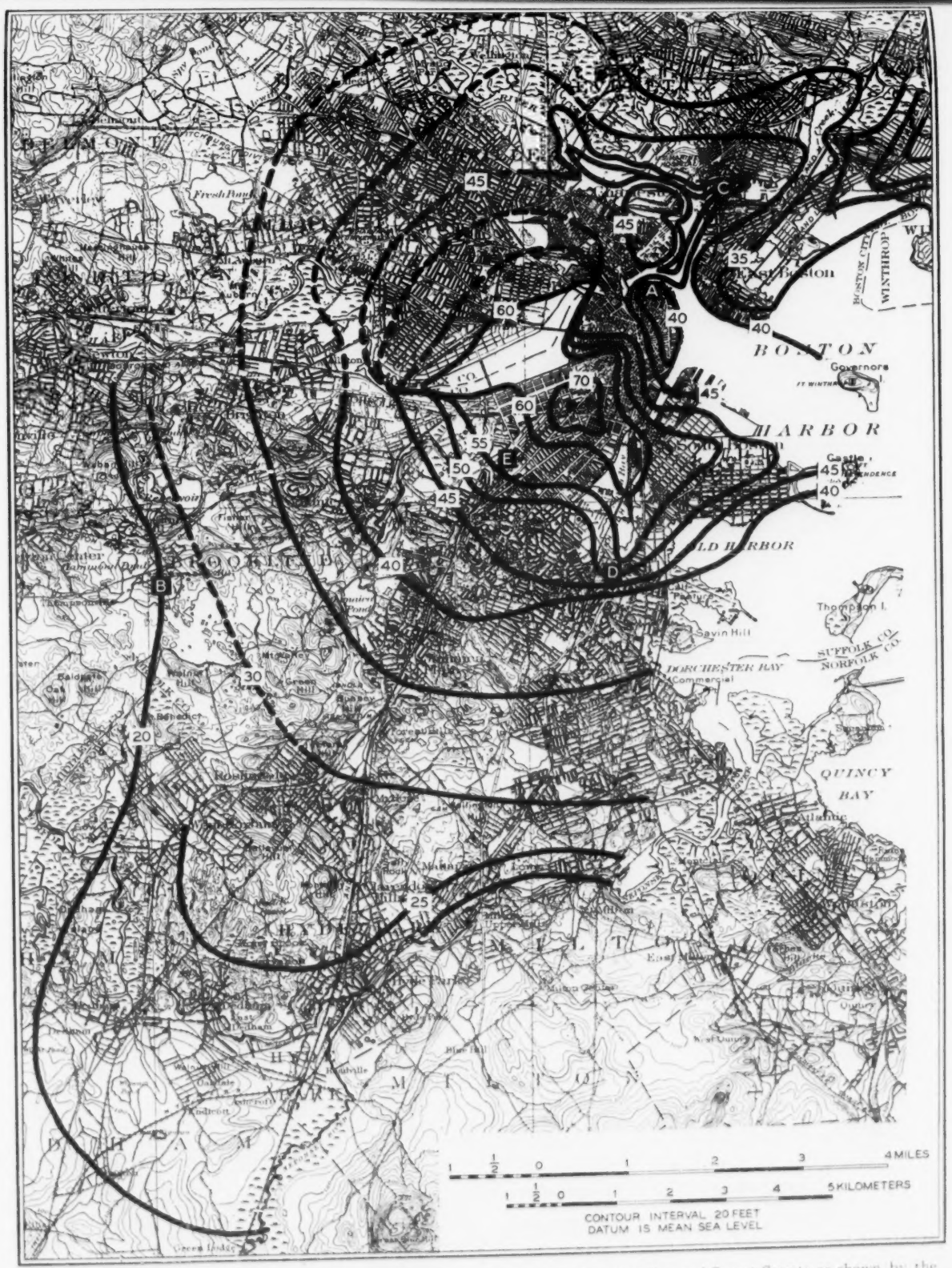


Fig. 8—Field strength contour map of Boston, Mass. Transmitter located at Berkeley and Stuart Streets as shown by the solid circle. Frequency = 34.6 mc. Field strength in decibels above one microvolt per meter with one ampere in a half wave antenna.

more or less unobstructed optical path existed along Columbus Avenue.

While a detailed analysis of the attenuation of ultra-short waves over paths as complicated as those considered in this paper is impossible, the known facts indicate several general characteristics that seem worthy of mention for further experimental investigation. The fact that the field was found to be approximately proportional to the height of the antenna above the roof level of the surrounding buildings (in Fig. 4), rather than to the height above the ground, as in the case of transmission over level terrain free from buildings, confirms the expectation that the "ground" conditions in the immediate vicinity of the fixed terminal would play an important part in determining the magnitude of the received field strength.

It is perhaps correct to assume that for the fixed terminal the height to be substituted in equation (5) should be the height above the roof rather than the height above ground. This would reduce the "level terrain" curve of Fig. 2 by 10 db.

Figures 6, 7 and 9 show marked wave interference patterns which indicate reflections from a multiplicity of points in the immediate vicinity of the mobile terminal. Besides these variations in the magnitudes of the fields observed at points in close proximity to each other, which are undoubtedly caused by reflections from irregularities in the immediate vicinity of the terminal, there are the variations represented by the spread of the points about the mean curve (Fig. 2). That these variations may be attributable to conditions local to the terminal is indicated by the fact that the increase in field on the far side of salt water and the decrease in the field on the far side of hills, etc., do not persist at further distances. Even if the irregularities of the contours of Fig. 8 were removed the contours would not be concentric circles about the transmitter. At this stage in the development it would be unwise to attempt to say how much of the deviation of the contours from circles may be attributable to directional characteristics at the fixed terminal and how much may be due to the intervening terrain. Statistically speaking, however, it is safe to say that the additional attenuation attributable to deviations from level terrain such as is produced by the presence of buildings, is independent of the length of the transmission path, so that there is no *absorption* in the usual meaning of the word. Another experimental result that points to the possibility that the effect of the buildings is local may be deduced from the data presented by Jones [10] and by Trevor and Carter.[16] The latter showed that the more distant points lie on the inverse-square-of-distance curve expected for transmission over level terrain free from buildings. The nearby points, however, lie below this curve, indicating that the major effect of the buildings is a local one.

The greatest difficulty encountered in attempting to apply the results of these experiments to the pre-determination of the field to be expected for transmission in other cities would be the interpretation of the method of assigning values to the heights in equation (5). While sufficient data are not available to establish an empirical relation, for a first estimation it seems reasonable that if the height of the fixed antenna were measured from the general roof level of the surrounding buildings and the height of the mobile antenna were measured from the street level, the resulting field strengths would lie within the range of expected values.[19]

Field Strength Measurements

Receiver at a Fixed Location

The field strength data obtained with the receiver at Seaverns Street and the transmitter in the truck are shown in Fig. 10. These

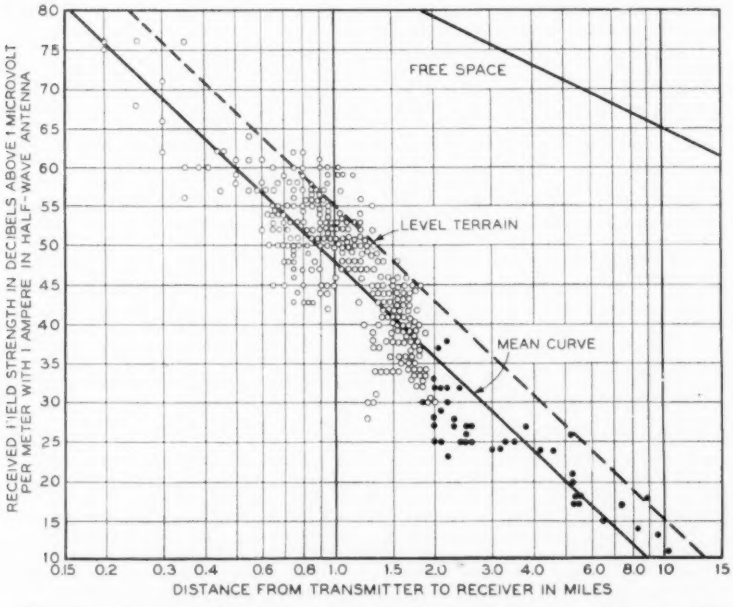


Fig. 10—Mass plot of field intensities measured with the transmitter at various distances about the Seaverns Street location. The open circles represent fields averaged over one-tenth mile intervals while the solid circles represent averages over one-half mile intervals.

[19] This should not be expected to be true for antennas close to the roof, since the field-strength obviously would not go to zero when the antenna height goes through the roof level. Extreme caution should be used in applying these results to cities in which the density and uniformity of buildings differ from those of Boston. When the topography departs greatly from that of level terrain, it would be difficult to infer the transmission conditions from the data here presented.

data may be represented also by an inverse-square-of-distance curve. By comparison with Fig. 2 no differences that can be attributed to the change in position of the fixed terminal nor the direction of transmission is evident. There is a small difference in the separation between the mean curves and the "level terrain" curves, but in both instances the mean curve lies very close to the level terrain curve (not shown) that would result from measuring the antenna height of the fixed terminal from the average roof level instead of from the ground.

Most of the nearby points were taken in the park system. They indicate that it is not more difficult to transmit through wooded areas than through built-up sections.

A field strength contour map illustrating the results obtained with the receiver fixed at Seaverns Street is shown in Fig. 11. The disturbing effect of hills is illustrated at several points on the map. There was a reduction of field when the transmitter was behind either Bussey Hill at *A* on the map or Green Hill at *B*, while the field was higher when the transmitter was between them. There was a rather deep minimum when the transmitter was immediately behind Parker Hill at *C* to the north, but half a mile farther away there was no noticeable effect.

Effect of Obstacles

In the course of the measurements some qualitative observations were made which will be summarized in this section. It was noticed in particular that when the receiver passed underneath an intersection of overhead trolley wires, the field was somewhat reduced. An example of this effect was observed at the intersection of Massachusetts and Huntington Avenues where the reduction was 15 db. At this point the maze of overhead trolley and support wires apparently constituted a fairly efficient screen for these waves. Observations made during the tests showed that sometimes the field was considerably reduced on the far side of hills as has been brought out by the contour maps. A large reduction in field strength resulted upon going behind a low hill on Saratoga Street on Breeds Island. The most striking example of this effect occurred with the measuring set at Seaverns Street and the mobile transmitter being driven from Huntington Avenue onto South Huntington Avenue (*C* of Fig. 11). Soon after rounding the corner at the foot of Parker Hill, which is 200 feet high, the average field dropped about 15 db.

The field was 20 db lower under Funeral Bridge than on either side of it. This is a stone and earth bridge appearing as a short tunnel to the road beneath it. The field was usually reduced upon passing underneath bridges of this general type of construction.

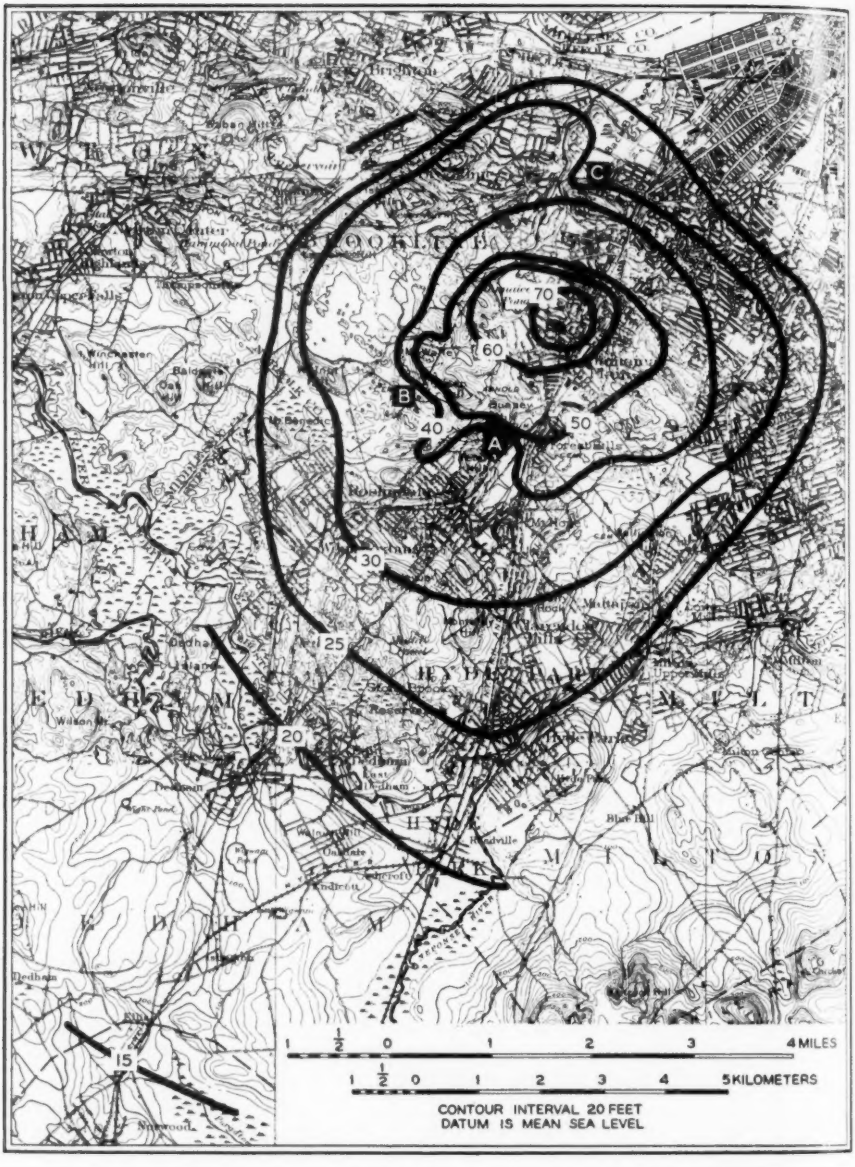


Fig. 11—Field strength contour map of Boston, Mass. Receiver located on Seaverns Street near Center Street as shown by the solid circle. Frequency = 34.6 mc. Field strength in decibels above one microvolt per meter.

A separation of the field strengths into those obtaining in the high-building area (indicated by open circles in Fig. 2) from those in the lower-building area (indicated by solid dots) shows that the attenuation is somewhat greater in the former.

No effect of the elevated railway structures on the average field was observed.

Noise Measurements

For reception in the car, by far the greatest interference is that caused by the electrical systems of passing automobiles. Special tests to determine whether or not street cars produced any noise gave a negative result. That is, under conditions where automobile noise was a limitation to reception, trolley noise was unaudible. While no special tests were made of the noise from elevated trains, at no time was it found objectionable.

With the receiver on top of the building at Berkeley and Stuart Streets, the predominating interfering noise was caused by an electrical substation next door. When the antenna was lowered approximately to the roof level the noise from the elevator motors and switching equipment in the pent house near by was well above any other noise. Upon raising the antenna to its proper position the elevator noise was reduced to a negligible amount compared with the power station noise, because of the combined effect of the directivity of the antenna and increased distance. The resulting noise was of approximately the same magnitude, indicating that the elevator switching noise was reduced by a fairly large factor in raising the antenna. This fact has an important bearing on reception of signals on the roofs of office buildings, since elevator switching noise is in general the limiting factor. Occasionally an automobile, started in the street below, would produce measurable interference. At Seaverns Street, however, the most objectionable noise was caused by the ignition systems of automobiles which were accelerating in low gear in the street below.

Acknowledgment

The authors wish to acknowledge the coöperation of the New England Telephone and Telegraph Company and the Graybar Electric Company, and also Commissioner E. A. Hultman and Signal Director T. A. J. Hayes, both of the Boston Police Department.

APPENDIX

Methods of Calibrating

In order to obtain an absolute calibration of the measuring equipment, the field strength at a distant point in space, at the same height

above the ground as the center of the receiving antenna, was obtained by a standard method of field strength measurement.[20] This equipment then was removed and the truck placed in such a way that its antenna was always at the point where the field strength had been determined. By pivoting the truck about this point the horizontal polar diagram of the mobile receiving equipment was obtained. It is shown in Fig. 12. As the field was known at this point in the absence

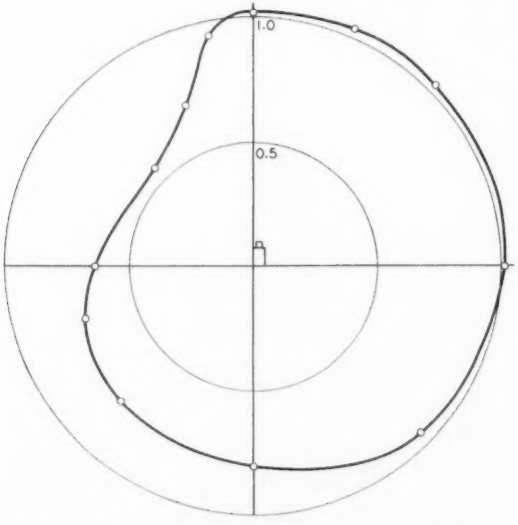


Fig. 12—Directional characteristic in the horizontal plane of the mobile receiver.

of the truck the attenuator settings gave a calibration of the receiving equipment for all directions. With this calibration as a standard, a comparison field generator which employed a loop radiator attached to the opposite side of the truck was calibrated also. The comparison field generator then was used throughout the test to check the calibration of the receiver. Since the polar diagram was fairly constant on the right side of the truck, care always was taken to orient the truck in such a way that the direction of the transmitter was at the right side rather than the left.

[20] This method consists of comparing the unknown field with a known field produced by a "Standard Field Generator," which is a small compact self-contained oscillator. It is very carefully shielded except for a small balanced loop extending in a vertical plane above the shield. A thermomilliammeter is located in the loop at the point of low potential with respect to the shield. From the reading of the meter and the dimensions of the loop, the field at nearby points may be computed. See *Proc. I. R. E.*, Vol. 21, pp. 430–431, March, 1933.

When the receiver was at Berkeley and Stuart Streets it was possible to employ the usual method of calibration [18] because the roof was flat. With the receiver at the Seaverns Street location, however, the gabled roof made it impractical to support the standard field generator opposite the midpoint of the antenna. In the latter case, accordingly, the constancy of the gain of the receiving equipment was depended upon in reducing the measurements to field strengths in absolute units. This lack of calibration did not introduce a large uncertainty, since the receiving equipment has been used over a period of years during which time its gain has remained constant within a few decibels.

Two-Way Tests

At the conclusion of this survey, actual two-way tests were made between a cruising car and fixed locations. For this purpose a car was equipped by E. B. Ferrell and R. C. Shaw [21] with an ultra-short-

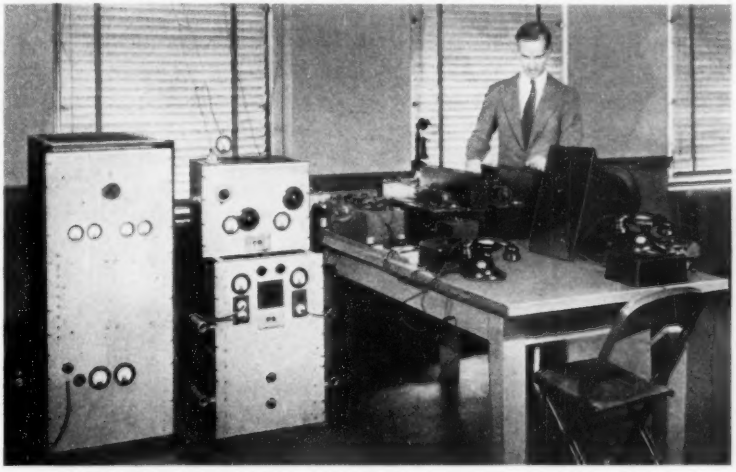


Fig. 13—Fixed terminal transmitting equipment located at Berkeley Street.

wave transmitter and receiver arranged for simultaneous two-way communication. A distinctive feature of this equipment was the use of a single antenna for simultaneously transmitting and receiving. This was made possible by the use of a suppressor circuit in the receiver to prevent overloading of the first detector by the outgoing signal. With this suppressor circuit, which consisted of only a half

[21] Both of Bell Telephone Laboratories, Deal, New Jersey.

section of a simple band-elimination filter, it was possible to transmit and receive simultaneously on frequencies differing by only five percent.

The equipment at the fixed transmitting location is shown in Fig. 13.

The car was used for communication at distances up to about 5 miles from the fixed transmitter and up to a little over 3 miles from the fixed receiver. The circuit from the moving car to the fixed receiver was consistently good for distances up to about two miles.

An Application of Number Theory to the Splicing of Telephone Cables *

By H. P. LAWTHER, JR.

The consideration of a simple and practical splicing scheme for minimizing the recurrence of same-layer adjacencies among telephone circuits in long cables leads to a problem in Number Theory whose solution calls for some extension of the previous work in this field. The solutions for numbers not greater than 139 have been computed, and a table of these is included.

SOME time ago in connection with the placing of a long telephone cable the writer had occasion to attempt the specification of a splicing scheme designed to minimize the recurrence of same-layer adjacencies among the telephone circuits as they threaded their way through successive lengths of the completed cable. The task, superficially so simple, proved to be one of most intriguing difficulty, and the pursuit of the solution led a confused investigator stumbling into the province of number theory. That speculation upon an art so mundane as that of telephone cable splicing should have led to a proposition in the oldest and most neglected branch of mathematics seemed to be especially worthy of note, for few applications so practical have been found. In the course of the investigation certain small ground apparently was covered for the first time. It was felt, therefore, that the story would be of passing interest alike to the mathematician and to the engineer.

The present standard cables for long distance telephone service are manufactured as a series of concentric layers of conductor units contained within a cylindrical sheath. The conductor units are either pairs of quads of wires. The layers are one unit in thickness, and successive layers either spiral in opposite directions of rotation, or in the same direction but with different pitches. The feature of importance to this discussion is that in an unbroken length of cable any one conductor unit will experience shoulder-to-shoulder adjacency throughout this distance with the two conductor units lying on either side in the same layer, and its experience with these two conductor units will be unique. Cables usually are manufactured in uniform lengths of from 750 to 1000 feet, and a longer cable is made up from a succession of such

* Published in *Amer. Math. Monthly*, February, 1935.

lengths spliced end-to-end. At each splice point a large number of different splices is possible among conductor units. In general, wire-to-wire splices are not made, and considerable mixing up is achieved. For reasons which need not be given here it is considered desirable from the standpoint of crosstalk control that each telephone circuit experience the minimum amount of same-layer adjacency with every other telephone circuit.

For the purposes of this discussion it will suffice at present to consider the cross-section of a cable as a simple closed sequence of N consecutively adjacent units. As an example, the array presented by a circular picket fence would be of this character. Each conductor unit in a cable is identifiable, and it will be assumed that each has been "tagged" with one of the numbers 1, 2, 3, 4, $\cdots$, N in such sequence that units bearing consecutive numbers lie adjacent—remembering that unit No. 1 and unit No. N also lie adjacent. While this simple picture of the cable cross-section is representative truly of only a single layer structure, still the results of a study of it may be fitted to apply to practical cases. Schemes for accomplishing this will suggest themselved to the practical worker, and their discussion here would burden this presentation unduly.

Consider now two consecutive lengths in a completed cable and focus attention upon a conductor unit in one of these. At the splice point this conductor unit may connect to any one of the conductor units in the second length, and the two conductor units which lie alongside the latter in the same layer in the second length may connect to any two of the $N - 1$ remaining conductor units in the first length. As an extended conductor unit traverses the completed cable, then, it may experience same-layer adjacency successively with any possible combinations two at a time of the other extended conductor units, and in any order, sequence, or repetition of these as determined by the splicing scheme that is used. Since there can be but $[(N - 1)/2]$* totally different combinations two at a time of $N - 1$ different objects it is evident that $[(N - 1)/2]$ successive cable lengths is the maximum possible number for an extended conductor unit to traverse without incurring repetition of at least one of the same-layer adjacencies that occurred in the first of these lengths.

Any splicing scheme that is devised for practical use must embody the utmost in simplicity. For this reason it is considered highly desirable (1) that the required results be achieved through repetition of the same splicing instruction at consecutive splice points, and (2)

* The symbol $[x/y]$ means the greatest integer not greater than x/y.

that this instruction follow the simplest possible system—e.g., any two adjacent conductor units in one length of cable shall connect to two conductor units having a constant separation in count in the next length. The exposition which follows makes no attempt to solve the general problem, and seeks only to establish the results which can be realized when the above two simplifying restrictions are imposed. At the conclusion is added a description of a minor and acceptable deviation from the second restriction which will enable the practical worker to supplement these results and achieve the maximum possibilities in a number of cases sufficient for his needs. The problem now will be formulated.

Fig. 1	Fig. 2	Fig. 3
1 → 1	1 → 1	1 → 1
2 → 2	2 → 3	2 → 4
3 → 3	3 → 5	3 → 7
4 → 4	4 → 7	4 → 10
5 → 5	5 → 9	5 → 2
6 → 6	6 → 11	6 → 5
7 → 7	7 → 2	7 → 8
8 → 8	8 → 4	8 → 11
9 → 9	9 → 6	9 → 3
10 → 10	10 → 8	10 → 6
11 → 11	11 → 10	11 → 9

The three tabulations exhibited in Figs. 1, 2, and 3 show possible ways of splicing two pieces of eleven-unit cable together in systematic fashion. The left-hand columns indicate the consecutively adjacent conductor units in the first or reference piece of cable (remembering that No. 1 and No. 11 are adjacent), and the numbers opposite in the right hand columns indicate the conductor units in the second piece of cable to which splice is made. No importance attaches to the splicing of unit No. 1 to unit No. 1 in each instance. This is simply one of eleven possible "starts," and from the point of view of this discussion there is no preference among these. Note that with Fig. 1 two conductor units which lie adjacent in the first piece of cable connect to conductor units separated by a count of one (adjacent) in the second piece. With Fig. 2 conductor units which lie adjacent in the first piece connect to conductor units separated by a count of two in the second piece. With Fig. 3 conductor units which lie adjacent in the first piece connect to conductor units separated by a count of three in the second piece. Splices made in accordance with the schemes of Figs. 1, 2, or 3

will be described as made with a "spread of one," a "spread of two," of a "spread of three," respectively. It is readily shown that for a spread number s to be applicable to cable of N units it is necessary and sufficient that s be prime relative to N.

Figure 4 shows the splicing of six pieces of eleven-unit cable through

$$
\begin{array}{cccccc}
1 \rightarrow & 1 \rightarrow & 1 \rightarrow & 1 \rightarrow & 1 \rightarrow & 1 \\
2 \rightarrow & 3 \rightarrow & 5 \rightarrow & 9 \rightarrow & 6 \rightarrow & 11 \\
3 \rightarrow & 5 \rightarrow & 9 \rightarrow & 6 \rightarrow & 11 \rightarrow & 10 \\
4 \rightarrow & 7 \rightarrow & 2 \rightarrow & 3 \rightarrow & 5 \rightarrow & 9 \\
5 \rightarrow & 9 \rightarrow & 6 \rightarrow & 11 \rightarrow & 10 \rightarrow & 8 \\
6 \rightarrow & 11 \rightarrow & 10 \rightarrow & 8 \rightarrow & 4 \rightarrow & 7 \\
7 \rightarrow & 2 \rightarrow & 3 \rightarrow & 5 \rightarrow & 9 \rightarrow & 6 \\
8 \rightarrow & 4 \rightarrow & 7 \rightarrow & 2 \rightarrow & 3 \rightarrow & 5 \\
9 \rightarrow & 6 \rightarrow & 11 \rightarrow & 10 \rightarrow & 8 \rightarrow & 4 \\
10 \rightarrow & 8 \rightarrow & 4 \rightarrow & 7 \rightarrow & 2 \rightarrow & 3 \\
11 \rightarrow & 10 \rightarrow & 8 \rightarrow & 4 \rightarrow & 7 \rightarrow & 2
\end{array}
$$

Fig. 4

the successive application of five consecutive identical splices, each with a spread of two. Following the "key" of the first and second columns, the succeeding columns are written down immediately. Scrutiny of the sequences of numbers appearing in the several columns reveals at once the fundamental properties of the spread. For a cable of N units these are:

1. Successive applications of a spread of s for n times result in a spread of s^n.

2. A spread of minus s is equivalent effectively to a spread of plus s.

3. A spread of $KN + s$ (K is an integer: positive, negative, or zero) is the same effectively as a spread of s.

The problem of achieving the minimum possible recurrence of same-layer adjacencies among conductor units through the application of successive similar splices in accordance with a simple spread now may be stated formally in the terminology and symbols of number theory. If N, an integer, is the number of conductor units in the cable, and if s, an integer prime to N, is the spread number used, then it is required to find a value for s for which the companion relations

$$
\begin{aligned}
s^d &\equiv \pm 1 \pmod{N}, \\
s^b &\not\equiv \pm 1 \pmod{N}, \quad b < d
\end{aligned}
$$

determine the largest possible integer d.

From the foregoing introductory discussion it should be noted that

values for N less than 5 are of no significance to this problem. In the analysis which follows, therefore, no particular effort has been made to render the general conclusions capable of extension to these extreme and trivial cases.

It is necessary at this point to recall and introduce certain working material. First, there is the established theorem that every positive integer N greater than unity can be represented in one and only one way in the form

$$N = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t},$$

where $p_1, p_2, \cdots, p_t$ are different primes and $\alpha_1, \alpha_2, \cdots, \alpha_t$ are positive integers. Then there is the familiar number theory function $\phi(N)$ which indicates the number of positive integers not greater than N and prime to N.* If p is a prime number and α is a positive integer, then

$$\phi(p^\alpha) = p^{\alpha-1}(p-1);$$

also

$$\phi(p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}) = \phi(p_1^{\alpha_1}) \cdot \phi(p_2^{\alpha_2}) \cdot \cdots \cdot \phi(p_t^{\alpha_t}),$$

where $p_1, p_2, \cdots, p_t$ are different primes.

Then there is the λ-function defined in terms of the ϕ-function as follows:

$$\lambda(2^\alpha) = \phi(2^\alpha) \text{ for } \alpha = 0, 1, 2,$$

$$\lambda(2^\alpha) = \frac{\phi(2^\alpha)}{2} \text{ for } \alpha > 2,$$

$$\lambda(p^\alpha) = \phi(p^\alpha) \text{ for } p \text{ an odd prime,}$$

$$\lambda(2^{\alpha_1} p_2^{\alpha_2} p_3^{\alpha_3} \cdots p_t^{\alpha_t}) = M,$$

where M is the least common multiple of

$$\lambda(2^{\alpha_1}), \lambda(p_2^{\alpha_2}), \lambda(p_3^{\alpha_3}), \cdots, \lambda(p_t^{\alpha_t}),$$

$2, p_2, p_3, \cdots, p_t$ being different primes.† Finally, it is established that for two relatively prime integers s and N the value $\lambda(N)$ is the largest possible for the exponent m for which the relations

$$s^m = 1 \pmod{N},$$

$$s^n \not\equiv 1 \pmod{N}, \quad n < m,$$

* Euler, "Novi Comm. Ac. Petrop.," 1760–61, p. 74. Carmichael, "The Theory of Numbers," John Wiley & Sons, Inc., 1914, pp. 30–32. Dickson, "Introduction to the Theory of Numbers," Univ. of Chicago Press, 1929, Chap. I.

† Cauchy, *Comptes Rendus*, Paris, 1841, pp. 824–845. Carmichael, p. 53.

will hold, and that a value for s belonging to this exponent does exist.*

Here it is convenient to consider separately numbers of the two classes—those for which $\lambda(N) = \phi(N)$ and those for which $\lambda(N) < \phi(N)$. For numbers of the first class established theorems may be drawn upon to furnish a complete analysis. For numbers of the second class, however, it will be necessary to extend a bit beyond the ground covered by previous workers, and the steps will be given in considerable detail. This procedure coupled with the inherent complexity will render the treatment for the latter class much less compact and elegant than that for the former.

Case I. $\lambda(N) = \phi(N)$.

From the defined relation between the ϕ-function and the λ-function it follows that numbers of the class such that $\lambda(N) = \phi(N)$ are confined to the values

$$1,\ 2,\ 4,\ p^{\alpha}, \quad \text{and} \quad 2p^{\alpha},$$

where p is an odd prime and α is a positive integer.† For a number N of this class it is established that there exists a set of $\phi(\phi(N))$ numbers r, such that

$$r^{\lambda(N)} \equiv 1 \pmod{N} \quad \text{and} \tag{1}$$

$$r^{n} \not\equiv 1 \pmod{N}, \quad n < \lambda(N), \quad \lambda(N) = \phi(N). \tag{2}$$

Such a number is known as a "primitive root" of N.‡ From the properties of the primitive root r of the number N as defined by relations (1), (2) it follows readily that

$$r^{\lambda(N)/2} \equiv -1 \pmod{N}, \tag{3}$$

$$r^{n} \not\equiv \pm 1 \pmod{N}, \quad 0 < n < \frac{\lambda(N)}{2}, \quad \frac{\lambda(N)}{2} < n < \lambda(N). \tag{4}$$

First there will be considered the companion relations

$$s^{d} \equiv -1 \pmod{N}, \tag{5}$$

$$s^{b} \not\equiv \pm 1 \pmod{N}, \quad b < d, \tag{6}$$

and, from comparison with relations (3) and (4), these clearly are satisfied for s a primitive root of N and for $d = \lambda(N)/2$. That no

* Carmichael, p. 54 and pp. 61–63.
† Carmichael, p. 71.
‡ Gauss, "Disquisitiones Arithmeticae," Art. 52–55. Carmichael, pp. 65–71. Dickson, Sec. 17.

value for d greater than $\lambda(N)/2$ is possible is evident immediately. For let a be any integer prime to N. Then for some exponent k

$$r^k \equiv a \pmod{N} \quad \text{and}$$

$$a^{\lambda(N)/2} \equiv r^{k\lambda(N)/2} \equiv \pm 1 \pmod{N}.$$

Next there will be considered the companion relations

$$s^d \equiv 1 \pmod{N}, \tag{7}$$

$$s^b \not\equiv \pm 1 \pmod{N}, \quad b < d. \tag{8}$$

The reasoning just above shows that d cannot be greater than $\lambda(N)/2$. Suppose for the moment that d has this greatest possible value $\lambda(N)/2$. Relations (7), (8) then become

$$s^{\lambda(N)/2} \equiv 1 \pmod{N},$$

$$s^b \not\equiv \pm 1 \pmod{N}, \quad b = 1, 2, 3, \cdots, \lambda(N)/2 - 1.$$

These relations may be written

$$(s^{1/2})^{\lambda(N)} \equiv 1 \pmod{N}, \tag{9}$$

$$(s^{1/2})^{2b} \not\equiv \pm 1 \pmod{N}, \quad 2b = 2, 4, 6, \cdots, \lambda(N) - 2. \tag{10}$$

Now relations (9), (10), will be compatible with relations (1), (2), (3), (4) only if $\lambda(N)/2$ is an odd number, for otherwise the restrictions of relation (10) applying to the even numbered exponents from 2 to $\lambda(N) - 2$ inclusive would be in conflict with relation (3). For $\lambda(N)/2$ an odd number, then, relations (9), (10), are satisfield for $s^{1/2}$ a primitive root of N. Consequently, with relations (7), (8), $\lambda(N)/2$ is the largest possible value for the exponent d, and a value for s equal to the square of a primitive root of N permits this to be attained.

Case II. $\lambda(N) < \phi(N)$.

The inquiry for this case will be divided into four parts. In general $N = p_1^{a_1} p_2^{a_2} p_3^{a_3} \cdots p_t^{a_t}$ where $p_1, p_2, p_3, \cdots, p_t$ are different primes.

(a) First will be considered the case where $p_1, p_2, p_3, \cdots, p_t$ are all odd primes. Then $\lambda(N)$ is the least common multiple of $\lambda(p_1^{a_1})$, $\lambda(p_2^{a_2})$, $\lambda(p_3^{a_3})$, $\cdots$, $\lambda(p_t^{a_t})$. Suppose now that the highest power of 2 dividing any of the λ's divides $\lambda(p_i^{a_i})$. If this same power of 2 divides more than one of the λ's, arbitrarily select $\lambda(p_i^{a_i})$ as one of them. Then this power of 2 will be exactly that occurring in $\lambda(N)$. Now arbitrarily select p_j as any one of the odd primes other than p_i. Then clearly

$\lambda(N)$ will also be the least common multiple of

$$\lambda(p_1^{\alpha_1}),\ \lambda(p_2^{\alpha_2}),\ \cdots,\ \lambda(p_{j-1}^{\alpha_{j-1}}),\ \frac{\lambda(p_j^{\alpha_j})}{2},\ \lambda(p_{j+1}^{\alpha_{j+1}}),\ \cdots,\ \lambda(p_t^{\alpha_t}).$$

Now take

$$\begin{aligned} r &\equiv g_j^2 \pmod{p_j^{\alpha_j}}, \quad g_j \text{ a primitive root of } p_j^{\alpha_j}, \\ &\equiv g_k \pmod{p_k^{\alpha_k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha_k}, \\ &\qquad\qquad k = 1, 2, 3, \cdots, j-1, j+1, \cdots, t. \end{aligned}$$

The r thus chosen must be prime to each of the prime factors of N, and hence must be prime to N. Consequently it is known that

$$r^{\lambda(N)} \equiv 1 \pmod{N}.$$

Suppose that m is the smallest exponent for which the congruence

$$r^m \equiv 1 \pmod{N}$$

is true. Then it is noted that the chosen r is such that m must be a multiple of

$$\lambda(p_1^{\alpha_1}),\ \lambda(p_2^{\alpha_2}),\ \cdots,\ \lambda(p_{j-1}^{\alpha_{j-1}}),\ \frac{\lambda(p_{1j}^{\alpha_j})}{2},\ \lambda(p_{j+1}^{\alpha_{j+1}}),\ \cdots,\ \lambda(p_t^{\alpha_t}),$$

and the least multiple common to these is, of course, $\lambda(N)$. Therefore it can be written that

$$\begin{aligned} r^{\lambda(N)} &\equiv 1 \pmod{N}, \\ r^b &\not\equiv 1 \pmod{N}, \quad b < \lambda(N). \end{aligned}$$

Now suppose that for some exponent n less than $\lambda(N)$

$$r^n \equiv -1 \pmod{N}.$$

Then

$$r^{2n} \equiv 1 \pmod{N},$$

and if n is less than $\lambda(N)$, $2n$ is less than $2\lambda(N)$ and can only be equal to $\lambda(N)$. It would necessarily follow then that

$$r^{\lambda(N)/2} \equiv -1 \pmod{N},$$

and it would follow in turn that

$$r^{\lambda(N)/2} \equiv -1 \pmod{p_i^{\alpha_j}}.$$

However, r has been chosen such that

$$r^{\lambda(N)/2} \equiv (g_j^2)^{\lambda(N)/2} \equiv g_j^{\lambda(N)} \equiv 1 \pmod{p_j^{\alpha j}}.$$

This last relation is incompatible with the one immediately above, and it must be concluded that the assumption

$$r^n \equiv -1 \pmod{N}, \quad n < \lambda(N)$$

is false, and that for the r that has been chosen

$$r^{\lambda(N)} \equiv 1 \pmod{N},$$
$$r^b \not\equiv \pm 1 \pmod{N}, \quad b < \lambda(N)$$

and no exponent greater than $\lambda(N)$ is possible.

(b) Next will be considered the case where $p_1 = 2$, $\alpha_1 = 1$, and $p_2, p_3, \cdots, p_t$ are all odd primes. Select p_i as above and take p_j different from 2. Then take

$$\begin{aligned} r &\equiv 1 \pmod 2, \\ &\equiv g_j^2 \pmod{p_j^{\alpha j}}, \quad g_k \text{ a primitive root of } p_j^{\alpha j}, \\ &\equiv g_k \pmod{p_k^{\alpha k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha k}, \\ &\qquad k = 2, 3, 4, \cdots, j-1, j+1, \cdots, t, \end{aligned}$$

and the same line of reasoning may be repeated and the same conclusions reached as under part (a) above.

(c) Next will be considered the case where $p_1 = 2$, $\alpha_1 = 2$ and $p_2, p_3, \cdots, p_t$ are all odd primes. Since $\lambda(2^2) = 2$ take p_j different from 2, and for simplicity take p_j as 2. Then take

$$\begin{aligned} r &\equiv 1 \pmod 4, \\ &\equiv g_k \pmod{p_k^{\alpha k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha k}, \\ &\qquad k = 2, 3, 4, \cdots, t \end{aligned}$$

and the same line of reasoning may be repeated and the same conclusions reached as under part (a) above.

(d) Finally will be considered the case where $p_1 = 2$, $\alpha_1 > 2$, and $p_2, p_3, \cdots, p_t$ are all odd primes. Now 5 has the property that

$$5^{\lambda(2^{\alpha 1})} \equiv 1 \pmod{2^{\alpha 1}},$$
$$5^b \not\equiv \pm 1 \pmod{2^{\alpha 1}}, \quad \alpha_1 > 2, \quad b < \lambda(2^{\alpha 1}).$$

So by taking

$$r \equiv 5 \pmod{2^{\alpha_1}},$$
$$\equiv g_k \pmod{p_k^{\alpha_k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha_k},$$
$$k = 2, 3, 4, \cdots, t$$

it is concluded immediately that

$$r^{\lambda(N)} \equiv 1 \pmod{N},$$
$$r^b \not\equiv \pm 1 \pmod{N}, \quad b < \lambda(N).$$

The preceding formal analysis for Case I and Case II may be summed up as having established the following general theorem:

If N is a given positive integer and if s is an integer prime to N, then the largest possible exponent d for which the companion congruencial relations

$$s^d \equiv \pm 1 \pmod{N},$$
$$s^b \not\equiv \pm 1 \pmod{N}, \quad b < d$$

will be true is $\lambda(N)/2$ for numbers such that $\lambda(N) = \phi(N)$ and is $\lambda(N)$ for numbers such that $\lambda(N) < \phi(N)$, and a value for s belonging to this exponent in each instance does exist.

In order to apply the foregoing results to a practical case Table I has been prepared. In the left-hand column appear the numbers 5 to 139, inclusive. In the next column is listed for each number the value of $\lambda(N)/2$ or of $\lambda(N)$, depending upon whether $\lambda(N) = \phi(N)$ or $\lambda(N) < \phi(N)$. In the final column there is listed for each number a suitable value for the spread. There appears to be no advantage of one spread figure over another, and the listing of additional acceptable values is omitted in the interest of economy of space. For the numbers for which $\lambda(N) = \phi(N)$ and for which $\lambda(N)/2$ is odd care has been taken that the listed spread figures are primitive roots, and not the squares of primitive roots which were shown to be equally acceptable. This fact will be recalled later.

It was shown earlier that $[(N-1)/2]$ successive cable lengths would be the maximum possible number for an extended conductor unit to traverse without incurring repetition of at least one of the same-layer adjacencies which occurred in the first of these lengths. On referring to Table I it is seen that only for the prime numbers is this maximum attainable. The prime numbers are distinguished by the fact that for them $\lambda(N)/2 = (N-1)/2$, and each has been indicated by an asterisk. The composite numbers are seen to yield quite inferior results in general.

TABLE I

For each N there is listed the value d and a value s for which the companion relations
$s^d \equiv \pm 1 \pmod{N}$ $s^b \not\equiv \pm 1 \pmod{N}$, $b < d$
determine the largest possible integer d.

N	d	s	N	d	s	N	d	s
5*	2	2	50	10	3	95	36	2
6	1	5	51	16	5	96	8	5
7*	3	3	52	12	7	97*	48	5
8	2	3	53*	26	2	98	21	3
9	3	2	54	9	5	99	30	5
10	2	3	55	20	2	100	20	3
11*	5	2	56	6	3	101*	50	2
12	2	5	57	18	5	102	16	5
13*	6	2	58	14	3	103*	51	5
14	3	3	59*	29	2	104	12	7
15	4	2	60	4	7	105	12	2
16	4	3	61*	30	2	106	26	3
17*	8	3	62	15	3	107*	53	2
18	3	5	63	6	2	108	18	5
19*	9	2	64	16	3	109*	54	6
20	4	3	65	12	3	110	20	3
21	6	2	66	10	5	111	36	2
22	5	7	67*	33	2	112	12	3
23*	11	5	68	16	3	113*	56	3
24	2	5	69	22	2	114	18	5
25	10	2	70	12	3	115	44	2
26	6	7	71*	35	7	116	28	3
27	9	2	72	6	5	117	12	2
28	6	5	73*	36	11	118	29	11
29*	14	2	74	18	5	119	48	3
30	4	7	75	20	2	120	4	7
31*	15	3	76	18	21	121	55	2
32	8	3	77	30	2	122	30	7
33	10	5	78	12	7	123	40	7
34	8	3	79*	39	3	124	30	7
35	12	2	80	4	3	125	50	2
36	6	5	81	27	2	126	6	11
37*	18	2	82	20	7	127*	63	3
38	9	3	83*	41	2	128	32	3
39	12	2	84	6	5	129	42	14
40	4	3	85	16	3	130	12	3
41*	20	6	86	21	3	131*	65	2
42	6	11	87	28	2	132	10	5
43*	21	3	88	10	3	133	18	2
44	10	3	89*	44	3	134	33	7
45	12	2	90	12	7	135	36	2
46	11	5	91	12	2	136	16	3
47*	23	6	92	22	3	137*	68	3
48	4	5	93	30	13	138	22	7
49	21	5	94	23	5	139*	69	2

* The asterisk indicates a prime number.

For the benefit of the practical worker there must be described a slight deviation from the second simplifying restriction imposed at the beginning which will permit the maximum possibility to be realized if N is one plus a prime number. This artifice is based upon the fact

that for r a primitive root of a number N for which $\lambda(N) = \phi(N)$ and in particular for a prime number N

$$r^{\lambda(N)/2} \equiv -1 \pmod{N}.$$

This means that $\lambda(N)/2$ consecutive splices with a spread r result in a spread of minus one. It is readily shown that this in turn means that there will be two conductor units No. b and No. $b + 1$ in the first length of cable which ultimately will be extended to connect respectively to units No. $b + 1$ and No. b. In Fig. 4 two units No. 6 and No. 7 meet this requirement. To illustrate the use of this artifice it will be supposed that a cable of 12 units is to be spliced. Referring to Fig. 4. for guidance, the arrangement shown in Fig. 5 is set up readily. The first two columns indicate the splicing assignment, and the succeeding columns are then derived from these. The eleven units 1, 2, $\cdots$, 5, 6, 8, 9, $\cdots$, 11, 12 are assigned exactly in conformity with the scheme of Fig. 4, ignoring the break in sequence between No. 6 and No. 8. Unit No. 7 is then simply spliced to itself throughout.

$$\begin{array}{cccccc}
1 \to & 1 \to & 1 \to & 1 \to & 1 \to & 1 \\
2 \to & 3 \to & 5 \to & 10 \to & 6 \to & 12 \\
3 \to & 5 \to & 10 \to & 6 \to & 12 \to & 11 \\
4 \to & 8 \to & 2 \to & 3 \to & 5 \to & 10 \\
5 \to & 10 \to & 6 \to & 12 \to & 11 \to & 9 \\
6 \to & 12 \to & 11 \to & 9 \to & 4 \to & 8 \\
7 \to & 7 \to & 7 \to & 7 \to & 7 \to & 7 \\
8 \to & 2 \to & 3 \to & 5 \to & 10 \to & 6 \\
9 \to & 4 \to & 8 \to & 2 \to & 3 \to & 5 \\
10 \to & 6 \to & 12 \to & 11 \to & 9 \to & 4 \\
11 \to & 9 \to & 4 \to & 8 \to & 2 \to & 3 \\
12 \to & 11 \to & 9 \to & 4 \to & 8 \to & 2
\end{array}$$

Fig. 5

Undoubtedly there are other equally acceptable artifices for extending further the practical scope of the simple results. The prime numbers and the prime numbers plus one constitute nearly fifty percent of all numbers in the range in which the practical worker is likely to be interested, however, and when it is borne in mind that normally he has latitude in his choice of N it is seen that the material here presented is adequate for his needs.

The writer is indebted to Dr. D. H. Lehmer for pertinent suggestions. The entire treatment for the case of numbers for which $\lambda(N) < \phi(N)$ follows a line of attack suggested by Mr. Marshall Hall, and but for his helpful interest this presentation would have been lacking in formal completeness.

Contemporary Advances in Physics, XXIX
The Nucleus, Fourth Part

By KARL K. DARROW

The earlier parts of this series have dealt with the charge, the mass, the stability or instability, and the liability to transmutation, of the atom-nucleus; this one deals with the two remaining properties which are ascribed to nuclei, to wit, magnetic moment and angular momentum. Since these exhibit themselves chiefly by influencing the spectra of the atoms to which the nuclei belong, the bulk of the article is concerned with various laws of atomic spectra; advantage is taken of the opportunity to describe some features of the electron-systems surrounding the nuclei, and to explain how the concept of the spinning electron enters into atomic physics. There follows an account of various experiments in which streams of atoms are deflected by inhomogeneous magnetic fields, and the laws of the deflections indicate the magnetic moment or the angular momentum of the nucleus or both. Finally there is a summary and tabulation of existing knowledge of these quantities.

The Nucleus as a Quantized Vector

UNDER this somewhat forbidding title I propose to discuss some phenomena—mostly spectroscopic, but in certain cases magnetic or even chemical—which are interpreted by supposing that the nuclei of atoms are endowed with two vectorial qualities, *magnetic moment* and *angular momentum*. One may say that these nuclei are to be visualized, no longer simply as particles possessed of mass and charge alone, but as bodies—usually, as congeries of particles both charged and uncharged—which are in incessant rotation: the spinning of the mass constitutes their angular momentum, the spinning of the charge a perpetual circular current-flow which is equivalent to a magnet. Why, then, should I not have entitled this section "The Spins and Magnetic Moments of Nuclei"? Chiefly because it might have suggested, at the very outset, that nuclear spins and magnetic moments are observed as clearly and measured as directly, as are nuclear masses and charges: which in the main is not so. With only a couple of exceptions (for magnetic moment) they are deduced from phenomena which certainly carry no obvious sign of their character. Indeed, what is called the nuclear angular momentum or the *spin* is distinguished by a feature, which is altogether strange and foreign to angular momentum of ordinary wheels and gyroscopes and the other spinning things of daily life; and it is by virtue of this foreign and paradoxical feature—and not any of the familiar qualities of spinning things—

that the concept of "spin" is a valuable addition to atom-models. This feature it is which is responsible for calling the nucleus a "quantized" vector. But I need not further justify at this point my suggestion that spin is a quality both strange and remote from immediate experience, for this will be only too evident in what is to follow.

It is desirable to make a longish excursion through some of those early-known and more-or-less familiar features of spectra, which are explained solely by calling into account the orbital or circum-nuclear electrons of the atom, without taking notice of any property of the nucleus excepting its charge which holds those electrons together. This excursion will have the incidental advantage of making us aware of the magnetic moment and the spin of the *electron*, though we cannot pause long enough to get an adequate appreciation of the scope and value of this concept of the "spinning electron." It is best to simplify these preliminary steps as much as is conveniently possible, and therefore I will speak of a single spectrum only; yet not the spectrum of hydrogen, for this would carry simplicity rather too far. By a queer paradox of which we are destined to see altogether too much, it often happens in quantum mechanics that the seemingly-simplest of all cases—the ones to which one would go by choice for a start—are liable to a singularly-confusing complexity quaintly known as a "degeneracy." Let us therefore take sodium for our example.

The sodium atom has a nuclear charge $+ 11e$, and eleven orbital electrons, of which the eleventh or "valence electron" is sharply contrasted with the other ten. Employing the original atom-model of Bohr, one visualizes the ten as describing a network of interlacing close-packed orbits enclosing the nucleus as in a sort of cage, while the eleventh darts about in some sort of a far-flung orbit which at one end may enter the cage, while the other end reaches far out into space ("penetrating orbits") or alternatively an entire orbit may encircle the cage completely, not entering it at all ("non-penetrating orbit").

We now consider the major feature of the sodium spectrum—which indeed is the only feature of the absorption-spectrum of tranquil sodium vapor—a beautiful converging series of lines. All of the lines correspond to transitions between one and the same state, *viz.* the normal state of the sodium atom, and the various members of a sequence of abnormal or excited states. These are distinguished by an index n to which consecutive integer values are attached; and all together they are known as the P sequence, to distinguish them from other sequences of states which a fuller study of the sodium spectrum (not confined to the absorption-spectrum of the tranquil vapor) discloses. The terminology, of course, is a detail. What is essential is,

that they are all discriminated from each other by something physical, and yet they all have something in common which distinguishes them as a whole from states of other sequences. As to the nature of these "somethings," the original theory of Bohr is perfectly explicit. All the states correspond to orbits of the valence electron; the different P states correspond to orbits of different but definite sizes and shapes; what all the P orbits have in common is, *a common value of the angular momentum of the electron revolving in its orbit.*

To repeat Bohr's argument here would be an unjustifiable use of time and space. Let me merely recall that when it was applied to the hydrogen atom, it led to several extraordinary agreements with the data of experiment, to which others have since been added. To some degree these are transferable to the sodium atom, especially since the electric field in which the valence-electron of the sodium atom mostly revolves is extremely like that in which the electron of the hydrogen atom always revolves (*i.e.*, beyond the "cage" to which I referred, the ten electrons of the cage effectively cancel the influence of the portion $+ 10e$ of the nuclear charge, leaving uncancelled only the field of a nuclear charge $+ e$ which is the same as the charge of a hydrogen nucleus). These agreements were the primary cause of the enormous role which angular momentum has ever since been playing in all atom-models. They were responsible also for the numerical values now assigned to angular momenta which figure in atoms; for according to the original theory, the orbital angular momentum of the P states is $2(h/2\pi)$, and all the other values which it may take for states of other sequences are other small-integer-multiples of $(h/2\pi)$; and while these values have since been somewhat altered without impairing the numerical agreements on which they rested and on which now the new ones rest, it remains true that *all angular momenta occurring in atom-models are expressed as multiples of* $\frac{1}{2}(h/2\pi)$, the multiplying factors being integers usually smaller than ten.

As my words have already implied, there are various types of angular momentum nowadays fitted into atom-models, the one already described—hereafter to be called the "orbital angular momentum"—being only one and the first. We turn now back to the principal series of sodium to discover why another type is required.

Examined with a sufficiently good spectroscope, each "line" of that series is found to be actually a close doublet. These imply that each of what I have been calling the P states is actually a pair of states.[1] (The confusion introduced into language by referring to one and the

[1] That it is not the normal state which is resolved into a pair is proved by various facts which it is not necessary to mention here.

same spectroscopic object sometimes as a line and sometimes as a pair or group of lines, and the corresponding confusion about states, are very bad impediments to clarity of exposition, but there is simply no way of getting around them.) Such is in fact the case. The two members of a pair have a common value of the index n and a common value of the orbital angular momentum of the valence-electron, and yet there must be something physical which distinguishes them. To leap at once to the conclusion: in the atom-model, this something is *orientation*.

But, orientation of what with respect to what? We have as yet introduced only one outstanding direction, only one vector, into the atom-model. The outstanding direction is that of the normal to the orbital plane of the valence-electron; the vector is the orbital angular momentum. To these it is necessary to supply a second vector,—as it turns out, a second angular momentum.

Since it may occur to some reader that the natural place to seek this angular momentum is among the electrons of the cage—that we should begin by assigning a net or resultant angular momentum to the ten electrons which we have hitherto so much neglected—I will recall that such was actually the first suggestion. It prevailed during the early twenties, and was generally accepted; but it suffered from certain disadvantages, which now there is no particular reason for retelling at length. Yet it was of the greatest assistance in preparing the ground and the technique for the suggestion which superseded it in the middle twenties, that the second angular momentum is to be ascribed to the electron itself; the electron is to possess, like the earth, not only a motion of revolution but also a motion of rotation. (As for the ten electrons of the cage, their angular momenta both of rotation and of revolution are so oriented as to balance one another out, and we have made no error in neglecting them.)

Now there are two vectors and two directions in the model of the sodium atom: that of the axis of the rotating electron, and that of the normal to the orbit—the spin momentum and the orbital momentum. To speak of different orientations of the one with respect to the other is now sensible. But different orientations must correspond to different energies if they are to explain the data, since the two lines of a principal-series doublet are separate and distinguishable because and only because the two members of a pair of P-states differ in energy. Why should they?

This happens to be the easiest question of the lot, or at any rate the one which can be answered from classical physics. The rotating electron is a magnet, by virtue of its whirling charge. Also it behaves

as though it were moving through a magnetic field, and this is not so easy to grasp, since we have postulated merely that it is revolving in the electrostatic field surrounding the nucleus and the cage of inner electrons. Yet if we were to shift our frame of reference and imagine, not the nucleus standing still and the valence-electron revolving around it, but the electron standing still and the nucleus revolving around *it*—then we should have no difficulty in realizing that the revolving nucleus, being a charge describing a closed path, and being therefore equivalent to a current, would produce a magnetic field. Reverting now to our original frame of reference, we may bring the magnetic field back with us, and say that the *spinning* electron revolves both in the electrostatic field aforesaid and in the magnetic field, and its energy is influenced by both. This is not a very sophisticated way of looking at the matter, and there is a tricky little relativistic detail in the shifting of the frame of reference, which produces an error of a factor 2 if disregarded; but it serves to bring out the idea. The energy of the valence-electron in its orbit is affected by this quasi-magnetic interaction due ultimately to the fact that it is a magnet moving through an electrostatic field; and *the value of the energy depends on the orientation*—on the angle θ between the axis of the magnet-electron and the normal to the orbital plane. One now sees readily that there will be a minimum energy occurring when these two directions are parallel, and a maximum energy occurring when they are anti-parallel. But why then do we not find the individual P-state spread out into a continuous band of states corresponding to all the energy-values between these two extremes, and all the infinity of different orientations between the value 0° and the value 180° of the angle θ?

This is no question which classical physics can solve. The fact that the individual P-state, or what would otherwise be the individual P-state, is split into two instead of into an infinity—this fact implies that only two orientations occur in nature, are "permitted," as the phrase is; and this instance of *quantization of direction*, like all the other instances of quantization, is a consequence of the quantum-mechanical constitution of the world. More lucid instances occur when the atom with its electrons is immersed in an applied magnetic field of known intensity; that is, when sodium vapor is exposed to the measurable field of a large-sized magnet, and its spectrum is observed. We will take up some of these instances before beginning with the assignment of spin to the nucleus.

When a magnetic field of moderate strength is applied to sodium vapor, each of the doublets of the principal series is split up into a pattern of several lines or "components." I can no longer say that

the P-states themselves are split up into as many components as are the lines: that was a happy coincidence while it lasted, but it does not repeat itself henceforward (nor usually). From the subdivision of the lines it is necessary to deduce the subdivision of the initial and the final states of the corresponding transitions: that is a classical task of spectroscopy, which we may assume to have been achieved. It is found that the normal state of the sodium atom (not belonging to the P sequence) is resolved by the magnetic field into two components, while of each pair of states belonging to the P sequence, one is resolved into two and the other into four components. Incidentally, the separations of these components are proportional to the strength of the magnetic field. It appears, therefore, that the sodium atom possesses properties of a magnet, quantized in direction; or rather, that in different states it is equivalent to different magnets, since in certain states it has two permitted orientations in the field, while in others it has four.

As might be guessed, the magnet to which the atom-as-a-whole is equivalent is a sort of *resultant* of the two magnets, spinning electron and "orbital magnet," which have already been separately inserted into the model. This quantum-mechanical resultant, however, possesses a couple of peculiarities, into which we shall have to look rather carefully. To lead up to them, it is desirable to look at the two special cases in which (*a*) there is no spin and (*b*) there is no orbital angular momentum, so that the resultant reduces to a magnet of one of the two types. Strictly speaking, case (*a*) never occurs in sodium, but to work it out is useful, nevertheless.

I have spoken of the electron revolving in its orbit as being equivalent to an "orbital magnet": now is the time to fortify that statement by giving a cardinal relation between the magnetic moment of that orbital magnet (not now the magnetic moment due to the electron-spin!) and the angular momentum of that orbital motion. It is sufficient to work out the relation for a circular orbit. Let r stand for the radius of the circle, v for the speed of the electron, $n(=v/2\pi r)$ for the number of times per second that the electron runs around the circle; e for the charge, m for the mass, p for the angular momentum of the electron, M for the magnetic moment of the system. The revolving electron is equivalent to a current[2] ne/c flowing in a circuit enclosing the area πr^2; the magnetic moment of such an affair is equal to the current-strength times the enclosed area:

$$M = (ne/c)\pi r^2, \tag{1}$$

[2] The factor c enters in because e is commonly expressed in electrostatic units, whereas in equation (1) the current must be expressed in electromagnetic units.

while for the angular momentum, we have:

$$p = mvr, \tag{2}$$

and eliminating vr between (1) and (2), we get:

$$M/p = e/2mc. \tag{3}$$

Here we have on the left the *ratio of magnetic moment to angular momentum*, a very important quantity for all these categories of atomic and subatomic magnets. It is equated to $e/2mc$, a value which is correct whenever we are dealing with the magnet constituted by the orbital motion (not the spin!) of an electron; this equation in fact is valid for any sort of an orbit described in a central field, and is one of the few that have survived unamended all of the stages in the evolution of Bohr's original theory into quantum mechanics. I will rewrite the equation thus:

$$M/p = g(e/2mc), \quad g = 1, \tag{4}$$

and this is meant to imply that for other categories of atomic and subatomic magnets, the ratio of the moments is not always equal to $e/2mc$, which is true. In general it is the custom to characterize any one of these magnets by giving its value of g. *Orbital magnets, then, are characterized by the value unity for the g-factor.*

We next ascertain the energy which the orbital magnet possesses by virtue of being in the applied field H. Letting α stand for the angle between the direction of the field and the axis of the magnet, we find for the torque exerted on the magnet by the field,

$$T = -MH \sin \alpha \tag{5}$$

and integrating to obtain the energy in question,

$$U = \int T d\alpha = MH \cos \alpha. \tag{6}$$

This we now write as follows:

$$U = (MH/p)p \cos \alpha = (geH/2mc)\ p \cos \alpha, \tag{7}$$

and here is as good an opportunity as any to recall a well-known theorem of classical mechanics, fundamental in the theory of the gyroscope. When a torque is acting upon a rotating body, the body precesses around the direction of the field responsible for the torque; and if the torque be equal to a constant T_0 times the sine of the angle between the field-direction and the axis of the rotating body, then the angular velocity ω of the precession is equal to the ratio between T_0

and the angular momentum of the rotation.[3] This ratio is the one enclosed in parentheses in equation (7): we may therefore write:

$$U = \omega p \cos \alpha. \tag{8}$$

The precession of which ω is the angular velocity is known as the *Larmor precession*. To keep this precession in mind as a feature of the atom-model is usually desirable, though not constantly necessary. It must be supposed to occur not only when an atom is immersed in an extraneous magnetic field, but also when two of the subatomic magnets within a single atom are influencing one another.

We have treated the (non-existent) extreme case in which there is nothing but the angular momentum and the magnetic moment of the orbital motion of a single electron to be taken account of; we turn now to the other extreme in which there is nothing but the spin of a single electron to be taken into account. Were we still confined to the original atom-model of Bohr, this case would be equally non-existent; for no electron-orbit could have a vanishing angular momentum unless it were a straight line passing to and fro through the nucleus, and this was formerly excluded as unthinkable. Quantum mechanics, however, assigns the value zero to the angular momentum of a valence-electron in a state of the S sequence (to which the normal state of the sodium atom belongs). Whether the student prefers to visualize a straight-line "orbit" for such a case, or a spherical cloud of charge or of "probability-of-charge," is to some extent a matter of taste, though usually the latter is the better policy. For such a state, there is no angular momentum and there is no magnetic moment save those of the electron-spin itself.

To this spin of the electron—whether isolated as in this extreme example, or compounded with an orbital motion into a resultant—we are compelled by various reasons to assign the value $2(e/2mc)$ for that important ratio of magnetic moment to angular momentum. Otherwise expressed: *the spin of the electron is characterized by the value 2 for the g-factor.*

There is a classical argument for this assertion, based on an evaluation of the ratio in question for a sphere of homogeneous charge rotating about an axis passing through its centre. There is a more powerful quantum-mechanical argument, based on the fact that when Schroedinger's fundamental equation of wave-mechanics was amended by Dirac to be conformable with relativity, there appeared in it a term attributable to a whirling charge with a g-factor of 2. Apart from

[3] Slater and Frank, "Introduction to Theoretical Physics," Chapter X.

this last, the strongest argument is furnished by the validity of the verifiable formula for the g-factors of various atoms in various states, which we shall presently be deriving. For, inasmuch as in nearly every state of an atom there are both electron-spins and electron-orbits, and the net magnetic moment and net angular momentum of the atom are sorts of resultants of these, the g-factor of the atom-as-a-whole varies from state to state and from one kind of atom to another in a remarkable fashion, which imposes a stringent test on the contemporary atom-model.

In explaining how this test is satisfied, I can no longer postpone specific numerical statements about the angular momenta of electron-orbits and electron-spins and the atoms into which these enter. These statements will be made confusing by the fact that to each one of these angular momenta it will be necessary to assign two different numbers. This is one of the impediments which are unavoidable in fitting visualizable atom-models to the results of quantum-mechanical theory, and which to avoid, some theorists would be willing to forego models altogether.

Were it not for this impediment, I could say quite simply that in every P-state of the sodium atom, the valence-electron has a spin, with angular momentum $\frac{1}{2}(h/2\pi)$, and is moving in an orbit with angular momentum $(h/2\pi)$; that as regards the two members of each of the aforesaid pairs of P-states, these two angular momenta are oriented parallel for one member and anti-parallel for the other, so that the net angular momentum of the atom-as-a-whole is $\frac{1}{2}(h/2\pi)$ in one case and $\frac{3}{2}(h/2\pi)$ in the other; that when an atom with a net angular momentum of $\frac{1}{2}(h/2\pi)$ is exposed to an applied magnetic field, it orients itself either parallel or anti-parallel to the field, so that the projection of its angular momentum upon the field-direction is either $+\frac{1}{2}(h/2\pi)$ or $-\frac{1}{2}(h/2\pi)$; that when an atom with a net angular momentum of $\frac{3}{2}(h/2\pi)$ is exposed to an applied magnetic field, it orients itself in one or another of four permitted ways so that the projection of its angular momentum upon the field-direction is either $+ 3/2$ or $+ 1/2$ or $- 1/2$ or $- 3/2$ times $h/2\pi$.

This sort of thing *is* frequently said in the literature, and one must realize its limitations. The trouble is, that quantum mechanics prescribes for these angular momenta (but not for their projections on the field-direction!) magnitudes which differ from those which I have been giving. For the spin s of the electron, it substitutes $\sqrt{1/2 \cdot 3/2}(h/2\pi)$ for $\frac{1}{2}(h/2\pi)$; for the orbital motion l in the present case, it substitues $\sqrt{1 \cdot 2}$ for the factor unity whereby $(h/2\pi)$ was multiplied; in the two values j' and j'' of angular momentum of the atom-as-a-

whole, it substitutes $\sqrt{1/2 \cdot 3/2}$ and $\sqrt{3/2 \cdot 5/2}$ for the factors $1/2$ and $3/2$. In order that these new values of j' and j'' may be resultants of s and l, it is necessary that the angular momenta of orbital motion and of spin should be neither quite parallel nor quite anti-parallel to one another—the two permitted values of the angle between them must differ from 0° and 180°; and this also is admitted by quantum mechanics. The last two clauses of the foregoing paragraph remain, however, unchanged.

Then why not say from the outset that the spin of the electron has the magnitude $\sqrt{1/2 \cdot 3/2}(h/2\pi)$, and that all the other angular momenta occurring inside an atom have the magnitudes assigned to them by quantum mechanics? Inertia of habit, sanctified by years of earlier theories, is itself a mighty obstacle: after so long a time of saying that the electron-spin is $1/2$, the world of physics could scarcely get accustomed to saying that really it is $\sqrt{1/2 \cdot 3/2}$ (in terms of $h/2\pi$ as unit). There is also a difficulty of printing to be considered: these numbers have often to be used as subscripts; it is bad enough to print $1/2$ or $3/2$ as a subscript, without resorting to their quantum-mechanical substitutes. The most serious reason, however, is, that the original $1/2$ and $3/2$ (and, of course, their analogues in the many other kinds of atomic states) lend themselves uniquely well to stating how many permitted orientations there are. Thus in the present example, I quoted 2 and 4 respectively as the number of permitted orientations in a magnetic field, of certain P states for which the angular momenta were designated as $j(h/2\pi)$ and j had the values $1/2$ and $3/2$ respectively. I was reminded of those numbers by the rule that they are equal to $(2j + 1)$. Had I kept in mind only the magnitudes $\sqrt{3/4}$ and $\sqrt{15/4}$ assigned by quantum mechanics to these angular momenta, the rule would not have been available.

One should therefore keep in mind both the "marker" or "quantum-number" of an angular momentum, of which the foregoing $1/2$ and 1 and $3/2$ are examples; and the numerical value assigned by quantum-mechanics to the magnitude of that angular momentum. Fortunately this is rendered easy by the fact that there is a general formula for the latter in terms of the former, with which we can now make acquaintance in the course of taking a deeper plunge into the lush notation of spectroscopy; as follows:

The angular momentum of electron-spin has the quantum-number s and the magnitude $\sqrt{s(s + 1)}$, and s is always equal to $1/2$.

The angular momentum of the orbital motion of an electron has the quantum-number l and the magnitude $\sqrt{l(l + 1)}$, and l may have the various values $0, 1, 2, 3 \cdots$.

The angular momentum of the atom-as-a-whole has the quantum-

number j and the magnitude $\sqrt{j(j+1)}$; for the sodium atom j may have the various values $1/2, 3/2, 5/2 \cdots$.

I next give the rule for compounding the last of these three angular momenta out of the first two, it being as I earlier said a "sort of resultant" thereof.

Rule for compounding l and s into j. If l is greater than s, start with the numerical value of $(l + s)$ and write down all the sequence of numbers spaced at unit intervals from $(l + s)$ to $(l - s)$ inclusive: to wit $(l + s)$, $(l + s - 1)$, $(l + s - 2)$, $\cdots$ $(l - s)$. These $(2s + 1)$ numbers are the permitted values of the quantum-number j. If on the other hand s is greater than l, start with the numerical value of $(s + l)$ and write down all the sequence of numbers spaced at unit intervals from $(s + l)$ to $(s - l)$ inclusive: to wit, $(s + l)$, $(s + l - 1)$, $(s + l - 2) \cdots (s - l)$. These $(2l - 1)$ numbers are the permitted values of the quantum-number j.

(It will be noticed that this rule is much more generally phrased than is required for the case of sodium, where $s = 1/2$, and it suffices to say that $j = 1/2$ for $l = 0$ and $j = l \pm 1/2$ for $l > 0$. If, however, we were dealing with an atom having more than one valence-electron, s might be replaced by a quantum-number different from $1/2$—not because the individual electrons would have new values of spin, but because the spins of two or more of them would be compounded—and the general phrasing of the rule would then be required). And now, to close (temporarily) the sequence of quantum-numbers and of rules:

Suppose the atom immersed in a magnetic field of strength H, parallel to the z-direction. It may then take any of several distinct permitted orientations, these being denoted by various values of a quantum-number m_j. In any such orientation the projection of the angular momentum of the atom-as-a-whole upon the z-direction or field-direction is equal to $m_j(h/2\pi)$. To ascertain how many permitted orientations there are and what are the corresponding projections, start with the numerical value of j and write down all the sequence of numbers spaced at unit intervals from $+ j$ to $- j$, inclusive: to wit, $j, j - 1, j - 2, \cdots - j$. These $(2j + 1)$ numbers are the permitted values of the quantum-number m_j.

We now start out upon a train of reasoning which leads to the remarkable verifiable formula already once alluded to, the successfulness of which speaks more powerfully than any other single test for the rightness of this elaborate hypothetical structure which has been devised for the atom.

Note, in the first place, that the rule given above for the permitted orientations of the atom in the applied magnetic field is in accordance

with a major fact of experience: the energy-values of the atom in its various "magnetic levels," as I shall hereafter call them, are known from spectroscopy to follow upon one another in a uniform evenly-spaced sequence. This is just what the rule requires: for if M denotes the magnetic moment of the atom and θ_j the angle which it makes with the z-direction, the energy due to the field is equal to $MH \cos \theta_j$, which is $MH(m_j/\sqrt{j(j+1)})$, and this changes by uniform steps as m is changed from each of its permitted values to the next.

Now recalling the importance of the ratio of magnetic moment to angular momentum, and noticing that $M/\sqrt{j(j+1)}$ is none other than this ratio, and introducing the g-factor of equation (7), and the precession ω of equation (8), we may write for ΔU the energy-difference between one magnetic level and the next:

$$\Delta U = g(e/2mc)H = \omega, \tag{9}$$

so that a measurement of the energy-difference or separation of two magnetic levels of an atom gives immediately the value of g for that atom. One sees at once how to make a special test of the value 2 assigned to the g-factor of the electron-spin; for when the sodium atom is in any state for which $l = 0$, it is the spinning valence-electron which contributes the whole of the magnetic moment and the angular momentum of the atom; and when from the spectrum of sodium vapor in the magnetic field the value of ΔU is determined for these states, it is precisely the value $2(e/2mc)H$ which is found.

To get a notion of what actually happens in the general case, it is best to take a sheet of paper and make a graphic composition of the angular momenta. These three—s, l, and j, to denote them by their quantum-numbers—are to be laid down as a triangle having sides of the lengths $\sqrt{s(s+1)}$, $\sqrt{l(l+1)}$, and $\sqrt{j(j+1)}$; for convenience I drop out the common factor $h/2\pi$ for the next few lines. The cosines of the three angles are obtained in terms of the sides by applying the well-known trigonometric formula and getting three equations of which here is one,

$$s(s+1) = l(l+1) + j(j+1) - 2\sqrt{l(l+1)}\sqrt{j(j+1)} \cos \theta_{l,\,j}. \tag{10}$$

The magnetic moment of the orbital electron-motion is a vector parallel to l and of the length $g(e/2mc)\sqrt{l(l+1)}$, with *unity* put as the value of g. The magnetic moment of the electron-spin is a vector parallel to s and of the length $g(e/2mc)\sqrt{s(s+1)}$, with *two* put as the value of g. Owing to the inequality of these g factors, the resultant of the magnetic moments is not parallel to j. We could easily calculate its magnitude and direction, but they are not relevant. It is

to be presumed that s and l are constantly revolving or precessing around the direction of j,—the triangle aforesaid is constantly revolving around its side j as a fixed axis, while retaining its size and shape unchanged. If we resolve the resultant angular momentum into a component parallel to j and another component perpendicular to j, the latter will be forever changing in direction and its average will vanish, leaving only the former as a perpetual constant. But this former component we can evaluate by projecting separately upon the j-direction the magnetic moments associated with orbital motion and spin, and adding the two projections. Thus for the magnetic moment of the atom-as-a-whole we get an effective average which is a vector parallel to the angular momentum of the atom-as-a-whole, and it is of the magnitude:

$$M_a = [\sqrt{l(l+1)}\cos\theta_{l,j} + 2\sqrt{s(s+1)}\cos\theta_{s,j}]\frac{e}{2mc}\frac{h}{2\pi} \qquad (11)$$

(where I have restored the factor $h/2\pi$). If we work this out with the aid of (10) and the similar equation for $\cos\theta_{s,j}$, and then divide it by the angular momentum $(h/2\pi)\sqrt{j(j+1)}$ and by $e/2mc$, we get *the g-factor for the atom-as-a-whole*—commonly denoted by g_j—in terms of the quantum-numbers s, l and j:

$$g_j = 1 + \frac{j(j+1) + s(s+1) - l(l+1)}{2j(j+1)}, \qquad (12)$$

and this is the celebrated *g-formula*, which is tested by applying magnetic fields H to atoms, splitting their stationary states into clusters of levels, measuring the separation between successive levels of a cluster, equating it to $\omega = g(e/2mc)H$, evaluating g and comparing it with the value which the right-hand member of (12) assumes when in it s, l and j are given the values appropriate to the state from which the cluster of levels was formed. So great is the variety of atomic states, so great the number of different triads of values of s, l, j represented among them, that the study of even a single element like sodium produces many different checks of the validity of (12); and since many different elements have been studied, the total of the available verifications of the g-formula, and therefore of the intricate network of its underlying ideas, is considerably impressive.

The temptation of going onward and onward into the details of these properties of the extranuclear electrons and their orbits is difficult to resist, but it must be overcome, for the field is practically endless. I must add only, that when an atom possesses two or more valence-

electrons, their orbital momenta $l_1, l_2 \cdots$ are likely[4] so to orient themselves as to form a fixed resultant L, and their spin-momenta s_1, $s_2 \cdots$ are likely to orient themselves so as to form a fixed resultant S; then L and S are likely so to orient themselves as to form a fixed resultant J, following in so doing the rule of page 295. Considered as quantum-numbers, J and S may have half-integer values only ($1/2$, $3/2 \cdots$) or full-integer values only ($0, 1, 2 \cdots$) according as the number of valence-electrons is odd or even, while L for any atom may have full-integer values only. The three resultants are vectors of magnitudes $\sqrt{J(J+1)}$, $\sqrt{S(S+1)}$ and $\sqrt{L(L+1)}$. The g-factor associated with L is unity and the g-factor associated with S is two, and the value of g for the atom-as-a-whole is given by (12) with capital letters replacing the small ones; so that the g-formula is verifiable with atoms of all kinds, as I intimated before. A final point: one might expect the conplexity to go on increasing tremendously from one end to the other of the Periodic Table, but there is a counteraction. In all atoms excepting the lightest, most of the electrons have oriented their orbits and their spins in such a way that they have interlocked themselves into groups or "closed shells" for which L is zero and S is zero and J is zero and the magnetic moment is zero, as have the ten electrons of the "cage" of the sodium atom to which I alluded. The so-called valence-electrons are those few which have not been locked into any such a cage. It is this quality which makes the Periodic Table periodic; but this must be left for some other place.

We arrive at last at the nuclear moment.

Suppose that even with these spins and these orbital motions of all the extra-nuclear electrons, we have not yet exhausted the internal angular momenta of the atom, and that the nucleus itself possesses one. Suppose, to be specific, that the nucleus has an angular momentum with a quantum-number I and a magnitude $\sqrt{I(I+1)}(h/2\pi)$, and a propensity for orienting itself in distinct permitted directions with respect to the other angular momenta of the atom. How shall we detect this, and how shall we determine I?

It is practically necessary to be yet more specific. One could probably not tell *a priori* whether the nuclear angular momentum would tend to orient itself with special respect to individual electron-momenta, or with special respect to some resultant or in particular to that grand resultant of all electron-momenta which we have denoted by J. However, in the cases which have been successfully analyzed, it

[4] This is the description of what is known as "Russell-Saunders coupling" or "*LS* coupling"; in certain states of certain atoms, the mutual orientations of the vectors conform to different schemes.

turns out that the last condition is the one which prevails. Suppose then that I compounds itself with J according to the following rule, repeated word-for-word with appropriate changes of symbol from the rule for compounding s and l into j:

Rule for compounding I and J into F. If I is greater than J, start with the numerical value of $(I + J)$ and write down the sequence of numbers spaced at unit intervals from $(I + J)$ to $(I - J)$ inclusive: to wit, $(I + J)$, $(I + J - 1)$, $(I + J - 2)$, $\cdots (I - J)$. These $(2J + 1)$ numbers are the permitted values of the quantum-number F. If J is greater than I, start with the numerical value of $(J + I)$ and write down the sequence of numbers spaced at unit intervals from $(J + I)$ to $(J - I)$ inclusive. These $(2I + 1)$ numbers are the permitted values of the quantum-number F.

The quantum-number F refers to a vector of magnitude $\sqrt{F(F + 1)}$ $(h/2\pi)$, which has taken over from J the role of the angular momentum of the atom-as-a-whole, being the resultant of J and of the nuclear angular momentum I.

Now if all these suppositions are correct, we may expect to find not individual states, but whole serried clusters of states, corresponding to individual values of J. If out of the manifold term-system of an atom we select a state for which $J = 1/2$, one for which $J = 3/2$, one for which $J = 5/2$ and so on as far upward as our knowledge extends, we may expect on close scrutiny to find that these apparent states are actually clusters, each cluster comprising a number of states which for one or two or more of the lowest values of J may be equal to $(2J + 1)$, but for higher values reaches and remains at a limit which we identify as $(2I + 1)$.

What is observed with the spectroscope, though, is not the individual state, but the line which reveals a transition between two different states. What in a feeble spectroscope appears as a single line, and is attributed to a transition between two states with resultant electronic angular momenta (I fear no shorter term will serve henceforth) j' and j'', should in an excellent spectroscope appear as a cluster of lines due to transitions between the several members of two clusters of states.

This again is exemplified by the principal series of sodium. As I said earlier, this appears in a feeble spectroscope as a series of single lines, each of which is resolved by a good spectroscope into a doublet. This structure, by the way, is called the "fine structure" of the lines; and this it is which indicates that the P-states of sodium are close pairs, and which thus invites and requires the introduction of the quantum-numbers j and s and the spin-momentum of the electron.

With a very good spectroscope indeed, each member of each principal-series doublet is in its turn resolved into a pair. This structure is called the "hyperfine structure" of the lines; and this it is which indicates that a still further subdivision of the states is necessary, and invites and requires the introduction of the quantum-numbers F and I and the angular momentum of the nucleus. Indeed, the concept of the nucleus as a quantized vector was invented or discovered (whichever word the reader may prefer) during the interpretation of hyperfine structure of spectrum lines.

Hyperfine structure of lines or states (for the name is applied to both) is usually more crowded and compact than fine structure, and yet there are exceptions: the fine structure of hydrogen is much harder to resolve than the hyperfine of (say) the familiar mercury lines 2537 and 5461, which itself was called fine before the theory was developed. These structures, however, are generally near and often, it is to be suspected, beyond the utmost capacities of the most refined of optical instruments; whence, in many cases, extraordinary difficulties in measuring or even estimating the separations, the relative intensities, actually the mere number of the distinct lines forming a hyperfine pattern; observers of great skill will often disagree with one another, and judgment will often depend on a photograph taken with a spectroscopic instrument such as an echelon or an etalon, which looks totally different from the pictures obtained with gratings or prisms. Perhaps this last is an advantage after all, as it discourages attempts by the inexpert to interpret published photographs. Often several different isotopes of an element produce different patterns which signify different values of I, and are so nearly superposed on one another as to make analysis superlatively hard. Hyperfine structure is for the present, and quite probably will be forever, the "last frontier" of spectroscopy.

The task of deriving, from the hyperfine line-pattern connecting two states or (better) state-clusters, the hyperfine subdivision of the state-clusters or "hyperfine multiplets" themselves, is again an example of the classical function of spectroscopy, which we shall take as having been achieved. Actually it involves, of course, the use of selection-principles, themselves connected with the atom-model, but omitted from this article in order not to complicate it still more. Some confusion may be prevented if I state that in our favorite case of sodium, where the fine-structure splitting of the principal-series lines implies a splitting of the P-states only, the hyperfine splitting implies something more complex: it is due jointly to hyperfine structures of both the P-states and the normal S-state, the latter being predominant.

We now consider briefly the methods of determining the quantum-number I of the nuclear angular momentum.

(*a*) The *ideal method* is the one already described: investigate line-clusters connecting state-clusters of as many different values of J as there are; ascertain thus the number N of states per cluster; verify that N is equal to $(2J + 1)$ whenever J is less than or equal to some particular value J_m (say), and that it is equal to $(2J_m + 1)$ whenever J is equal to or greater than J_m. All this being verified, the value of I must be J_m.

One seldom if ever finds such a programme as this worked out very fully. The difficulties seem to be that, at best, it is a lot of work to analyze the hyperfine-structure of even one line, let alone a great number; while at worst, lines connected with states of certain J-values, high ones especially, may be quite unobservable; also there is the striking fact that in a given spectrum a very few lines or even one alone may have their hyperfine-structures spread out so much more broadly than all the rest, that research is practically concentrated on them alone. (One hears so much about 4722 of bismuth as almost to have it blotted from mind that bismuth has other lines!) But of course if one is willing to accept without test the rule for compounding the vectors I and J, then it suffices to discover and analyze a state-cluster for which N is less than $(2J + 1)$.

(*b*) *The intervals between the members of a state-cluster* may give a clue to the value of I. These intervals are of course energy-differences, and the fact that they exist shows that there are forces between the spinning nucleus and the system of revolving and spinning electrons which surrounds it. If these forces are magnetic, then they may reasonably be expected to vary as the sine of the angle $\theta_{I,J}$ between the angular momenta of nucleus and extranuclear electron system; for either the magnetic moments of these two will be parallel to the respective angular momenta, or else (by the reasoning of page 297) their non-parallel components will presumably change so rapidly as to be ineffective, leaving only the parallel components to be detectable. Comparing the different orientations of I and J which correspond to the several values of F and thus to the states of the cluster, one sees that their energies—or rather, the parts W_F thereof which are due to the interaction—should then vary as $\cos \theta_{I,J}$: putting for which the formula based on (10)

$$W_F = \text{const.}\frac{F(F+1) - I(I+1) - J(J+1)}{2\sqrt{I(I+1)}\sqrt{J(J+1)}}. \qquad (13)$$

Insert in (13) the permitted values of F, which are $(2I + 1)$ or $(2J + 1)$ in number according as J is or is not greater than I, and are spaced at unit intervals from $(I + J)$ downwards (page 299); call them, in order of descending magnitude, F_m, F_{m-1}, F_{m-2} $\cdots$. Form the $2I$ or $2J$ consecutive first differences between the so-computed permitted values of W_F; call them ΔW_m, ΔW_{m-1}, ΔW_{m-2}, $\cdots$ Then as is readily worked out,

$$\Delta W_m : \Delta W_{m-1} : \Delta W_{m-2} \cdots :: F_m : F_{m-1} : F_{m-2} \cdots$$
$$:: (I + J) : (I + J - 1) : (I+J - 2) \cdots.$$

The successive energy-differences or *intervals* should stand to one another as the successive members of the chain of integers (or half-integers, as the case may be) stepped off at unit intervals and stretching from $(I + J)$ downwards.[5]

This is an *interval-rule* based on a specific notion of the intra-atomic forces (the sine-law aforesaid), and having analogues in the parts of atomic theory having to do with the interactions between electrons the extra-nuclear electrons only. If verified, it enables one to determine $(I + J)$ and therefore I from the analysis of a single cluster of states with a single value of J, even when J is smaller than I and the preceding method would fail. Much use has been made of this method, and there are a few cases in which a fairly accurate measurement of a chain of intervals has shown that it closely agrees with a chain of consecutive integers or half-integers, though more usually the intervals are small and the measurements rough and it is merely assumed that there is perfect agreement with that particular succession of half-integers or integers with which there is the nearest apparent agreement.

(*c*) *The relative intensities of the members of a line-cluster* are capable of giving information about the quantum-numbers of the states which they connect, provided one adapts quantum-mechanical formulae developed for transitions into which the nuclear angular momentum does not enter. The formulae are of appalling complexity, while intensity-measurements, especially when one is working so near the limits of the possible as when hyperfine-structure is being measured, are notoriously liable to error. This method is probably to be classified as by far the least reliable, for the present at any rate.

[5] The biggest interval may be that between the highest and the next-to-highest energy-value, or that between the lowest and the next-to-lowest; whichever case is realized gives a clue to the "sign" (page 318) of the magnetic moment; usually the former corresponds to a positive, the latter to a negative moment, but features of the extra-nuclear electron-system may cause this statement to be reversed. Incidentally, relative intensities of lines also have a bearing on the sign of the moment.

(*d*) *The phenomenon of alternating intensities in band-spectra* serves to reveal the spins of a few kinds of nuclei, and in a very interesting and reliable way, but must be left for another occasion.

There remain the methods which involve the use of a magnetic field in one way or another, and some of which incidentally tell most of what we know about the magnetic moments of nuclei, though nowhere near so amply or so exactly as we should like.

First it must be said that the analogy between the vectors I and J on the one hand, L and S on the other, which thus far has been so full and helpful, breaks down completely when the atom is exposed to a magnetic field of ordinary strength. Were the analogy perfect, an atom in a state distinguished by the quantum-number F for its total angular momentum would act as a rigid spinning body and would be able to assume $(2F + 1)$ discrete orientations in the magnetic field, corresponding to $(2F + 1)$ magnetic levels. This would be true of each of the $(2I + 1)$ or $(2J + 1)$ states comprised in what I have been calling a "cluster" with a common value of J, though the value of F and hence of $(2F + 1)$ would differ from one state to the next. The magnetic levels would be distributed in groups, each corresponding to a different value of F. The numbers in the different groups would be unequal. The total number for all the groups or states of the cluster would amount, as the reader can figure out, to the product $(2I + 1) \times (2J + 1)$.

It is altogether probable that this is precisely what does happen in magnetic fields so weak as not to separate the magnetic levels perceptibly (their separation being then, it will be recalled, proportional to the field-strength). Yet in fields strong enough to produce a measurable effect, the disposition of the magnetic levels has only one thing in common with this hypothetical distribution. Their total number *is* precisely $(2I + 1)\ (2J + 1)$. They are, however, distributed in $(2J + 1)$ groups, each consisting of $(2I + 1)$ levels; as though first of all the atoms were to forget their nuclear angular momentum and remember only their electronic angular momentum, and were to orient themselves in the field in the $(2J + 1)$ different ways which were prescribed for them (page 295) while the nucleus was still being neglected; and as though then they were to remember the nuclear angular momentum, and were to allow for it by adopting, in place of each separate one of the $(2J + 1)$ very different orientations, a group of $(2I + 1)$ orientations differing only a little from it and from each other.

This rather animistic idea is not very far from the model commonly conceived. It is supposed that in the strong magnetic field the nucleus

is somehow broken away from its interlocking with the system of extra-nuclear electrons; not in the sense that it is torn out of the system or that its electrostatic attraction for the electrons is suspended, but in the sense that somehow or other the rule for the compounding of I and J into various values of a resultant vector F is done away with. The system of electrons with its angular momentum J chooses among $(2J + 1)$ orientations with respect to the field, just as if the nucleus were not there; and the nucleus chooses among $(2I + 1)$ orientations with respect to the field, just as if the electronic system were not there. Nucleus and electrons, I and J, are said to be "decoupled" from each other; it is supposed that their angular momenta precess each at its own rate separately around the direction of the field.[6]

This account suggests that the energy-values of the magnetic levels would be given by the various values of the expression

$$M_e H \cos \theta_{J,\,H} + M_n H \cos \theta_{I,\,H}, \tag{14}$$

where $\theta_{J,\,H}$ and $\theta_{I,\,H}$ stand for the inclinations of the angular momenta J and I with respect to the field-direction, while M_e and M_n signify the magnetic moments of the extra-nuclear electron-system and of the nucleus, or, if these magnetic moments be not parallel to J and I respectively, then their projections upon the directions of J and I. The different groups of levels would then correspond to different permitted values of $\cos \theta_{J,\,H}$, the different levels of any one group to different permitted values of $\cos \theta_{I,\,H}$. The first term, one might say, would determine the $(2J + 1)$ separate energy-values which would occur if there were no angular momentum of the nucleus, while the second term would subdivide each of these into a group of $(2I + 1)$ levels.

It is found, however (as we shall later see) that the magnetic moments M_n of nuclei are always so very small by comparison with those of extranuclear electron-systems, that the second term of (14) is quite negligible. We have therefore to look for some other cause for the observable subdivision. This cause is thought to be the force between the moments of the nucleus and the electron-system. We assumed it to be overborne by the strong field in so far as its ability to control the quantized directions of the angular momenta is concerned,

[6] The analogy of I and J with L and S is restored when the impressed field is very strong, for then L and S are similarly decoupled from one another—"Paschen-Back effect" as distinguished from the "Zeeman effect" which we have hitherto been considering. Thus it is roughly correct to say that hyperfine structure reacts to a weak magnetic field as fine structure does to a strong one, though this statement should be carefully qualified if use were to be made of it.

but now must suppose it still to be potent enough to affect the energy of the atom. It adds a third term to (14) which is supposed to be of the form *const.* $M_e M_n \cos \theta_{I,\,H} \cos \theta_{J,\,H}$, but in any event will have $(2I + 1)$ distinct values for every value of J, and is not necessarily (or anyhow is not known to be) too small to account for the observed separations between the members of each group. It therefore allows us to estimate I by counting the members of such groups and equating their number to $(2I + 1)$, and this (when available) is one of the most acceptable ways of determining the angular momentum of the nucleus.

With a spectroscope one counts, as always, not the number of levels but the number of lines connecting them with some other family of levels, and expects that the two numbers will not be the same. By a rare if not unique coincidence, however, they *are* the same: the selection-principle which is involved is such, that each group of levels produces a group of an equal number of lines, or (in other words) if the influence of the nucleus resolves every state of the atom in a magnetic field into a group of $(2I + 1)$ different levels, then it also resolves every line connecting two such states into a group of $(2I + 1)$ different lines. There are in the literature magnificent photographs of the spectrum lines of bismuth exposed to a magnetic field, each line under high resolution exhibiting ten components and proving the value 9/2 for I.

It is, however, sometimes possible to count the levels directly, by sending a beam of fast-moving atoms through an inhomogeneous magnetic field which spreads it out into a diverging fan of smaller beams or pencils, each consisting exclusively of atoms having a certain distinctive value for the projection of the magnetic moment upon the field-direction. This requires a great refinement of the celebrated method of Gerlach and Stern, a refinement which has been achieved by Rabi and his school.

We take, as usual, sodium for our example. Consider a narrow beam of sodium atoms, moving with uniform speed along the x-direction into a region pervaded by a magnetic field which is parallel to the z-axis, and of which the magnitude H varies as rapidly as possible with z. Were it not for this variation of H with z, nothing would happen to the beam, for (to make the crudest possible picture) each atomic magnet would have both its north and its south pole exposed to the same field-strength, and one would be pushed as hard as the other was pulled, resulting in no net force upon the magnet and no deflection. But when the field varies with z and the atomic magnet is oriented otherwise than at right angles to the z-axis, the north and the south pole will be exposed to different field strengths, there will be a resulting

force and a resulting deflection of the flying atom. Force and deflection will increase with the z-component of the magnetic moment of the atom, and atoms with different values of this component will go in different directions.

First we disregard the influence of the sodium nucleus. The sodium atoms in experiments of this type are always in their normal state, for which I recall that $l = 0$ and $j = s$: the angular momentum and the magnetic moment are exclusively those of the spin of the valence-electron, the former having quantum-number $1/2$ and magnitude $\sqrt{\frac{1}{2}(\frac{1}{2}+1)}\ (h/2\pi)$, the latter being parallel to the former and having magnitude $g(e/2mc)$ times the magnitude of the angular momentum, with 2 for the value of g. Having this value, the angular momentum of the atom may orient itself in either of the two ways which are crudely (page 292) called "parallel" and "antiparallel" to the field, though it is better (page 293) to think of the two permitted inclinations to the field-direction as being $arc\ cos\ \frac{1}{2}/\sqrt{\frac{1}{2}(\frac{1}{2}+1)}$ and $arc\ cos\ (-\frac{1}{2}/\sqrt{\frac{1}{2}(\frac{1}{2}+1)})$. Half of the atoms are so oriented that the z-component of their magnetic moment is $\frac{1}{2}g(e/2mc)(h/2\pi)$, the other half so that the z-component has the negative of this value: the beam is split into two, diverging oppositely and symmetrically from the axis of x. The detection of this splitting is the Gerlach-Stern experiment.

Now we suppose that the sodium nucleus has an angular momentum of quantum-number I, as a result of which the two orientations aforesaid are not really two, but actually are two groups of $(2I + 1)$ not-very-different orientations apiece. The problem is, to refine the method sufficiently to bring out the fact (if it is a fact) that each of the two apparent beams aforesaid is actually a close group of several, and to count the several.

It is necessary to lengthen out the path of the atoms in the inhomogeneous magnetic field, since the longer their exposure to the deflecting agent lasts, the farther the separate beams are drawn apart; this means magnifying the scale of the apparatus and the volume which has to be kept evacuated, and carrying to a very high pitch the geometrical accuracy of its design, since the initial not-yet-separated beam must be exceedingly narrow and must be shot forth in a very-exactly-adjusted direction from its source into the field. It is essential also to reduce the broad distribution-in-velocity which the atoms owe to the fact that they come out of a furnace (in which sodium is being vaporized) with the random velocities of thermal agitation appropriate to the temperature of the furnace, and which would more than suffice to merge the beams which it is now desired to separate. One gathers that even at present it would not be possible to make the wished-for separation,

were it not for one more feature of the laws of the behavior of atoms and their internal angular momenta in the magnetic field.

I have described at length how, in magnetic fields of the field-strengths customary in spectroscopy, where the angular momenta I and J of nucleus and electron-system are "decoupled" and orient themselves independently in the field, each state with a given value of J is converted into $(2J + 1)$ groups of $(2I + 1)$ levels apiece. But in no field at all, I and J are "coupled" into a resultant F, or rather into one or another of several such resultants; and I mentioned that there is reason to suppose that, in fields very much weaker than the customary ones, this coupling subsists and the atom orients itself as a single entity in the field. I emphasized then (page 303) that different as are these two extremes, they have one feature in common: the total number of the levels corresponding to a single value of J, which at both extremes is $(2I + 1)\ (2J + 1)$. The practical usefulness of this theorem is diminished by the fact that some of these levels may have identical values of energy, but in the atom-model they are nevertheless distinct.

One naturally guesses that as the field-strength is increased from "very weak" to "customary," each level of the one extreme passes over into a level of the other extreme, so that for any field-strength low, intermediate or high there are always just $(2I + 1)\ (2J + 1)$ of them. There arise then the lesser problem of ascertaining the "correlation," i.e. which level of the one extreme goes over into which of the other; and the greater problem of ascertaining just how, for each of these continuously-definite levels, the energy-value and the component of the magnetic moment along the field-direction—which latter determines the deflection, and which let us call M_z—vary with the field-strength H. Formidable theoretical articles have been written on both of these problems, culminating in rules for the former and formulae for the latter. They were worked out originally for the behavior of the vectors L and S in applied magnetic fields, but are translated into rules and formulae available for our present interests by simply replacing these vectors with I and J and making corresponding changes in the g-factors.[7] For such an atom as hydrogen or sodium in its normal state, for which $J = \frac{1}{2}$, I will quote the formula from Breit and Rabi.

[7] Strictly one should take into account the influence of the magnetic field on the interrelations between L and S and on those between I and J simultaneously, but it usually happens that when H is increased to a magnitude which already suffices to decouple I and J pretty thoroughly, it is not yet great enought to do much to the coupling between L and S. I have spoken of this range of magnitudes as "customary," on the ground that it is usual in experiments on the Zeeman effect; but there is no good single word for qualifying it, inasmuch as it is simultaneously weak with respect to the (L, S) coupling and strong with respect to the (I, J) coupling.

We have seen that in the absence of magnetic field, the normal state is subdivided into two by the influence of the nuclear angular momentum I and its coupling with J. In one of these—call it N'—the vectors I and J are nearly parallel, and their resultant, the angular momentum F of the atom-as-a-whole, has the quantum number $(I + J)$ which is $(I + \frac{1}{2})$; in the weak field where the coupling is not broken, and the state N' maintains its identity, the number of permitted orientations is $(2F + 1)$ which is $(2I + 2)$. In the other—call it N''—the vectors I and J are nearly anti-parallel, and F has the quantum number $(I - \frac{1}{2})$; in the weak field the number of permitted orientations is $(2F + 1)$ which is $2I$. Adding, we get for the total number of magnetic levels the value $(4I + 2)$, which is equal to $(2I + 1)\ (2J + 1)$ as I stated. In the weak field, the different levels are distinguished by their values of the magnetic quantum-number m, which is defined by saying that the projection F_z of the angular momentum (of the atom-as-a-whole) on the field-direction is equal to $m(h/2\pi)$. The permitted values of m are $(I + \frac{1}{2})$, $(I - \frac{1}{2})$, $(I - \frac{3}{2}) \cdots$, $-(I - \frac{1}{2})$, $-(I + \frac{1}{2})$. The first and the last of these values are attached each to a single level, belonging (in the weak field) to the state N'; each of the others is attached to a pair of levels, one belonging to the state N' and the other to the state N''.

We know that each of these levels maintains its identity as the field-strength is increased, even when the coupling of I and J into F is broken down and the separate states N' and N'' lose their identities. We wish to know how the value of M_z for each level is varying as the field increases. Let a stand for $2m/(2I + 1)$; let b stand for the energy-difference between N' and N''; let g stand for the g-factor associated with the extra-nuclear electron-system and with the angular momentum J; let x stand for $(g/b)(eH/2mc)(h/2\pi)$. The formulae of Breit and Rabi are as follows[8]:

$$M_z = \pm \frac{a + x}{2(1 + 2ax + x^2)^{1/2}}\, g(e/2mc)(h/2\pi) \tag{16}$$

For the levels characterized by the extreme values of m(*viz.*, $\pm (I + \frac{1}{2})$) and initially belonging exclusively to N', the first factor is equal to one half and the two levels are distinguished by the two choices of sign, and M_z is independent of field-strength. With respect to the other values of m, the situation is more complex and curious. A single value of m, say $(I - \frac{1}{2})$, corresponds to two different values of M_z which are equal in magnitude and opposite in sign; the opposite value of m,

[8] Perhaps it is not superfluous to remark that in the factor $(e/2mc)$, the symbol m always stands for electron-mass, never for magnetic quantum-number.

in this case $-(I-\frac{1}{2})$, also corresponds to two different values of M_z which are equal and opposite, and each of which at zero field-strength (but not for $H=0$) coincides with one of the previous two. Now consider any two of these values of M_z which correspond to equal and opposite values of m and coincide with one another at $H=0$. As H is increased from zero, these two are altered in opposite senses, and one of them actually passes through zero and then reverses at its sign at a certain value of field-strength (m and hence a are negative, and M_z vanishes when $x=a$) while the other is shifted in the opposite sense and never vanishes (m and a are positive).[9] We shall presently see (page 310) that this behavior is the basis of one of the methods of evaluating I. One more peculiarity of these equations must be stressed: the magnetic moment of the nucleus nowhere appears in them! This becomes evident when the field-strength is put equal to zero and x vanishes, for then the several values of the right-hand member of (16) become simply the projections, upon the field-direction, of the magnetic moment of the extra-nuclear electron-system. Thus we have the paradox that in these experiments the magnetic field gives us information about the nucleus by virtue of the force which it exerts upon the atom, and yet this force is exerted practically upon the electrons alone, and not to any perceptible extent upon the nucleus.

The laws expressed in equation (16) have thus far assisted in three ways in the study of the nucleus:

First, in respect to the experiment which I was describing (page 306) when I began on this detour: as H is decreased from what I called the "customary" magnitude, the $(2I+1)$ levels constituting each of the there-mentioned groups draw gradually apart—i.e. they differ more and more in respect of the value of the component of the magnetic moment along the field-direction, which is what controls the deflection. The experiment must therefore be performed with field-strengths H which are sufficiently low, much lower than those customarily employed in the Gerlach-Stern experiment or in spectroscopy; and this is one of the distinctive features of the technique of Rabi and his school. Narrowness of the beam is all the more required, since dH/dz must be large enough to produce considerable deflection, and if both its value and the breadth of the beam in the z-direction were large, H could not be small in every part of the beam. The beam must also be made nearly homogeneous in speed, and this is done by a

[9] Exception must be made for pairs of values of M_z, both members of which correspond to $m=0$ and vanish at $H=0$; each member of such a pair departs farther and farther from zero, to equal extents in opposite senses, as H is increased from zero.

clever utilization of the law embodied in (16), for which I must refer to the original papers.—When the experiment was performed on sodium, it was found that the beam is split into eight components instead of merely two, proving the value 4 for the factor $(2I + 1)$ and the value 3/2 for the quantum-number of the nuclear angular momentum.

Second, suppose an experiment done by measuring the number of atoms which go through the deflecting field entirely undeflected. These are the atoms for which $M_z = 0$, and in ideal conditions there would never be any such atoms, excepting at one of the particular values of field-strength for which the M_z of one (or rather, two at a time) of the levels mentioned on page 309 is passing through zero; in actual conditions one would expect the curve of the number-of-undeflected-atoms *versus* the field-strength to exhibit peaks. On examining equation (16) one may see that there would be one peak for $I = 1$ or $3/2$, two for $I = 2$ or $5/2$, three for $I = 3$ or $7/2$, and so on.[10] The number of peaks by itself thus gives a partly ambiguous indication, but the ambiguity can be resolved by another theorem deducible from (16): if we denote by H_1, H_2, $\cdots$ the abscissæ of the consecutive peaks, then all the intervals $(H_i - H_{i-1})$ are equal in any case, but the value of H_1 is equal to the half or to the whole of their common value, according as I is a full integer or a half-integer.—The curve for caesium was found to display three peaks, and the second criterion showed that the value of I is a half-integer, therefore 7/2.

Third, when M_z is found by measuring the deflected beams in an apparatus where field and field-gradient are accurately known (not a stringent requirement in either of the two previous cases), one may use equation (16) to compute b: thus determining the "hyperfine-structure" separation between two states without an optical measurement! This has been done with both varieties of hydrogen, the heavy isotope and the light, because for these very important atoms the separation in question is far too small to be detected by any optical device: the method of magnetic deflection has proved itself the superior of the long-established arts of spectroscopy, hitherto regarded as the *ne plus ultra* of subtlety and refinement.[11] The results of these experiments are mostly quoted for their bearing on the magnetic moment

[10] Not counting the peak at $H = 0$ which (it is obvious) must always appear but has no bearing on the value of I.

[11] I should, however, perhaps make exception for the most delicate of these, which is the derivation of hyperfine structure from observations on the resonance radiation produced by polarized light acting on atoms of gases in magnetic fields, and is practiced by Ellett and his school at the University of Iowa. The complexity of the theory forbids a description of the method in this place, but several values of I have been obtained by it.

of the nucleus, that vector quality of which till now next to nothing has been said. Little indeed can be said about it with assurance, but we must consider at least that little.

I recall that the magnetic moment of the extra-nuclear electron-system is very simply ascertained by measuring the magnetic splitting-up of the stationary states, i.e. the energy-differences between the different orientations of an atom in an applied magnetic field, because it enters directly into the formula for those energy-differences; but although the nucleus itself produces a further splitting-up which in its turn is measured, the situation is so much more complicated that these measurements have no interpretable bearing on the value of the nuclear magnetic moment. For protium and for deuterium, the two isotopes of hydrogen, the magnetic moment of the nucleus has been directly measured by a magnetic-deflection method. For all the other kinds of atoms we are obliged to infer it by theory from the measured values of the energy-differences between the states of what I called a cluster, which are alike in respect of I and J and differ in respect of the mutual inclination of these vectors.

The theory can at least be illustrated by a quasi-classical derivation, though the differences between this and the quantum-mechanical method are not slight. One first visualizes the valence-electron as a charged particle running around and around its orbit, equivalent therefore to a steady current running around the orbit and producing a magnetic field at all points within the orbit and in particular at the point occupied by the nucleus; the nuclear magnetic moment is subjected to this field, and when it is shifted from one to another of its permitted orientations a certain amount of work must be done (or received) and constitutes the energy-difference in question. Supposing a circular orbit with radius r and angular momentum p, the argument commences like that of page 290; we have $pe/2\pi mr^2c$ for the strength of the equivalent current, pe/mr^3c for the field-strength which it produces at the centre of the circle where the nucleus is; we conceive the nucleus as having a magnetic moment M parallel to its angular momentum; we assign the quantum-number I to this angular momentum and the quantum-number l to that of the orbital motion of the electron, thus conceiving these as vectors having the magnitudes $\sqrt{I(I+1)}\ (h/2\pi)$ and $\sqrt{l(l+1)}\ (h/2\pi)$, which last is what I have been calling p. If we could ignore the spin of the electron, l could be replaced by J, and the torque exerted by the field upon the nucleus would be $M(pe/mr^3c)\sin\theta_{I,J}$. There would be two or more permitted values of $\theta_{I,J}$ corresponding to the various states of the cluster, and we should get the corresponding energy-values U by writing:

$$U = M(pe/mr^3c)\cos\theta_{I,J} = M(e/mc)(r^{-3})\sqrt{J(J+1)}h/2\pi\cos\theta_{I,J}, \tag{17}$$

and using expressions based on equation (13) for the cosine. Finally we should form the differences between the right-hand members of these equations, and equate them to the observed energy-differences between the states of the cluster, and solve for M.

Even this formula, when applied to the data, gives values of M of the same order of magnitude as do the more elaborate ones; otherwise there would be no point in quoting it. There is, however, very much to be done to improve it. There is the magnetic field produced at the nucleus by the spin of the electron. There is the alteration required by relativity. There is the task of applying quantum-mechanical rather than quasi-classical reasoning to the postulates. The procedure is strongly supported by the fact that it is copied from the argument which, in the theory of the interaction between the spin and the orbital angular momentum of the valence-electron, leads to a wonderful explanation of the fine-structure of the hydrogen spectrum. It is, however, certainly not perfect, since when applied to different states of a particular kind of atom it is likely to lead to different values of the nuclear magnetic moment, a result which either shows some of the mathematical methods of approximation to be faulty or else is a *reductio ad absurdum* of one or more of the postulates. The problem is in fact one of the great unmastered problems of atomic physics, and some believe that it is wrong to postulate that the nucleus can be regarded, in its interactions with the extra-nuclear electrons, as nothing but a simple magnet attached to a body having mass and charge. I shall therefore say nothing further about it, except for quoting the formula oftenest used in cases such as those of sodium and hydrogen, where the energy-difference b in question (to follow the notation of page 308) is that between the two members of a pair of states for both of which $L = 0$ and $J = S = 1/2$, while for one of them $F = I - J$ and for the other $F = I + J$:

$$b = (8\pi/3)\left(\frac{2I+1}{I}\right)M(eh/4\pi mc)\Psi^2(0), \tag{18}$$

the last symbol standing for the square of the value which the Schroedinger wave-function has at the nucleus, which is known exactly for hydrogen and approximately for other one-valence-electron atoms; the formula is due to Fermi. Applying this formula to the values of b for light and heavy hydrogen which they had ascertained by the magnetic-deflection method (page 310), Rabi Kellogg and Zacharias

got values for the magnetic moments of proton and deuteron [12] which are of the order of one one-thousandth of those associated with electrons. These values are stated (for a reason which may be obvious but will be set forth presently) as 3.25 and as 0.75 times the quantity $(eh/4\pi m_p c)$, where m_p stands for the mass of the proton; the uncertainty is given as 10 percent for the former, about 25 percent for the latter.

There remain the experiments whereby M was directly determined, for protons and for deuterons, from the force exerted by an inhomogeneous magnetic field on the nuclear moment itself. One cannot use the bare proton for such an experiment, since owing to its charge it would suffer, as it flies through the magnetic field, an electrodynamic force and a deflection by comparison with which the others would be trivial. One cannot use the isolated hydrogen atom, since just as in the case of sodium which we considered at such length, the force exerted by the field upon the magnetic moment of its electron would far outweigh that exerted on the nucleus. There remains the hydrogen molecule, which in its normal state has the convenient feature that the spins of its two electrons are oriented anti-parallel (in the loose sense of the term) and cancel one another out, while the angular momenta and the magnetic moments of their orbital motions likewise vanish. This seems to remove all the possible competitors to the nuclear moments, but there arises another which does not occur in individual atoms: the rotation of the molecule-as-a-whole, which has an angular momentum and a magnetic moment. This magnetic moment, however, is of the same order as those of the nuclei, and its contribution to the net magnetic moment of the nuclei can be estimated and subtracted from theirs. As for the nuclei, they may set themselves with their spins either parallel or antiparallel;[13] in the latter case their magnetic moments cancel one another, and observations on such molecules teach us only about the rotation, knowledge which is useful; in the other case their magnetic moments add, and the data of the experiment yield a value which is double the moment of the individual proton—or of the individual deuteron, according as the molecule is formed of two light atoms or two heavy atoms of hydrogen.

[12] This substitute for the names *deuton* and *diplon*, by which the nucleus of the H^2 atom (deuterium or "heavy hydrogen") has usually been known in America and England respectively, was recommended at a recent meeting of the American Physical Society by Dr. Urey, the discoverer of deuterium.

[13] This implies that nuclei conform to rules of quantization in direction relative to one another similar to those for electrons. This is true, and is superbly demonstrated by observations on band-spectra and by the chemical separation of the two kinds of hydrogen molecules ("ortho-hydrogen" and "para-hydrogen") here mentioned; but the story is much too long for this place.

The experiments were performed by Stern and his school. They are extremely delicate, owing to the excessively small value of the magnetic moment and hence of the deflection. Indeed, they might have been beyond instead of within the limits of the possible, were it not for two distinctive features of hydrogen: the just-mentioned doubling of the moment due to the parallelism of the spins of two nuclei, and the fact that a beam of slow-moving molecules can be utilized, because hydrogen does not have to be heated in a furnace in order to be vaporized, but will emerge as a molecular beam through a hole in the wall of a chamber no warmer, indeed very much colder, than the temperature of the room itself! For the magnetic moment of the proton, Frisch and Stern give 2.5 $(eh/4\pi m_p c)$ with an uncertainty of "at most 10 per cent"; for that of the deuteron, Estermann and Stern say that the value is between 0.5 and 1.0 times $(eh/4\pi m_p c)$.

We now turn to the ensemble of the estimates of nuclear angular momenta and of magnetic moments, and the laws and rules which they seem to obey.

Estimates of I have by now been made for some fifty-five elements; but this is not an adequate statement to make, for the nuclear angular momentum is one of those qualities—as I shall presently stress—which may vary from one isotope to another of a single element, and there are already several cases in which values of I have been reliably assigned to two or more different isotopes. The great majority are derived from optical analysis of hyperfine-structure; four or five from magnetic-deflection experiments; about ten from alternating intensities in band-spectra, most of these last being checked from the hyperfine structure of line-spectra.

The values are of very unequal merit, some being derived independently and concordantly from several different properties of hyperfine structure (pages 301–302), some being further sustained by deductions from band-spectra, while others are guesses based on a few rough observations of intensities or intervals. The unreliability of these last is of a type not to be described by giving a most-probable-value coupled with a probable error. A critical and analytical review of the lot by a neutral expert is badly needed.

The first thing which strikes the eye on viewing a tabulation is the immense preponderance of half-integer values of spin; but this is only one sign of the most important of the rules of nuclear momenta, which is one of the most important of the rules of nature. I recall that the "mass-number" A of an atom is the integer nearest to the value of its mass expressed in terms of one sixteenth the mass of the commonest oxygen atom as unit. The rule, then, is:

Atoms of odd mass-number have half-integer or "odd" spins; while among the atoms of even mass-number, two of the lightest have the full-integer or "even" spin of unity, while all the rest display no hyperfine-structure at all and thus probably have the "even" spin-value zero (though we must never forget the possibility that the lack of observable hyperfine-structure means that the magnetic moment is very small, the clusters of states therefore so compact that the spectroscope cannot resolve the hyperfine pattern, and the value of I therefore unascertainable).

Although not every kind of atom has yet been studied, the number of cases on which this rule is based is already so considerable that the discovery of an exception would be a sensation of the first order. Among the most striking exemplifications of the rule are those afforded by elements of many isotopes. Mercury, the outstanding example, displays a hyperfine-structure of wondrous complexity, which has been very successfully interpreted by assigning the value 3/2 of I to its odd isotope of mass-number 201 (Hg^{201}), the value 1/2 to its odd isotope Hg^{199}, and the value zero to its four principal even isotopes Hg^{198}, Hg^{200}, Hg^{202} and Hg^{204}; the lines attributed to these isotopes stand to one another in the intensity-ratios deduced from the relative abundances of the isotopes as measured by Aston. Cadmium has two odd isotopes for both of which $I = 1/2$, and several even ones exhibiting no hyperfine structure at all. Several elements have two isotopes, both of odd mass-number; in some of these cases both have the same value of I(e.g. gallium 3/2, rhenium 5/2), in other cases they differ (e.g. rubidium 3/2 and 5/2). The outstanding and very-certainly-known case of hydrogen is distinguished by the values 1/2 for the light and 1 for the heavy isotope.

(It may occur to the reader that if, say, four isotopes of mercury display no hyperfine-structure, then everything that I have heretofore said implies that their spectrum-lines and levels ought to coincide absolutely and be indistinguishable. These are, however, slightly separated, owing, it is presumed, to very slight differences in the fields surrounding nuclei of different isotopes even when they all belong to the same element and have the same nuclear charge and indistinguishable moments. The phenomenon, which is known as "isotope shift," is likely to be much studied in the theoretical physics of the near future).

As we go along the list of the atoms of odd mass-number from the lightest to the heaviest, the value 1/2 for I appears at the start; the value 3/2 at mass-number 6; 5/2 at $A = 35$; 7/2 at $A = 83$; 9/2 (the highest yet inferred) at $A = 93$. Further observations may change

these numbers somewhat, but it seems fairly assured that the higher values first appear later than the lower. Yet the lower and even the lowest appear repeatedly all through the list; for the five most massive atoms for which I is known, its values are 1/2 for Tl^{203}, Tl^{205}, and Pb^{207}, 9/2 for Bi^{209} and 3/2 for Pa^{231}. There is no sign of a periodic variation; the random fashion in which the display or the lack of detectable hyperfine structure are sprinkled among the elements would be sufficient proof that it is of nuclear origin, even were there no theory.

We recognize thus that I is a quality of nuclei which depends upon nuclear *mass* (not charge!) Moreover, everything connected with the concept of nuclear angular momentum—the behavior of the extranuclear momenta L and S and J of which it is an analogue, the nature of the phenomena which it is contrived to explain, the values of I deduced from these phenomena—all these things imply that I is a quantum-vector compounded out of the individual angular momenta of individual particles composing the nucleus, the quantum-numbers of these individual momenta being, like that of their resultant, integer multiples of 1/2. The very simplest model would consist, of course, of particles all having identical angular momenta of quantum-number 1/2.

Now there are two leading schemes for imagining a nucleus—let us say, of atomic number Z and mass-number A—as systems of minuter particles. According to the one, it consists of A protons and $(A - Z)$ negative electrons; according to the other, of Z protons and $(A - Z)$ neutrons.[14] It would be correct to say that the former was the leading scheme until about two years ago, the latter now; but as the facts about nuclear momenta and nuclear magnetic moments have had much to do with bringing on this change of favor, I will not take it for granted in advance.

If we take the nucleus to be a congeries of protons and negative electrons, then we *are* postulating a system of which all the particles are known to have spins of quantum-number 1/2—the simplest conceivable system, as I said; and we should still have in reserve the possibility of assigning orbital motions with orbital angular momenta to some or all of the particles, should it ever seem desirable. But if we take the nucleus to be a congeries of protons and neutrons, we are introducing particles of a kind for which the spin is unknown, and must be fixed by assumption. Since we can put 1/2 for the spin of the neutron, this affords little basis for choice.

[14] The fact that alpha-particles are often mentioned as constituents of nuclei is not in contradiction with this statement, since an alpha-particle is interpreted as 4 protons and 2 electrons by the one scheme, 2 protons and 2 neutrons by the other.

It is, however, an important point—a very important point—that the different schemes imply different numbers of particles for a given nucleus: $(2A - Z)$ by the former, and simply A by the latter. I have spoken of a notable rule contrasting the values of I for even mass-number with those for odd mass-number. Notice now that if we adopt the latter or proton-neutron scheme, the rule becomes: *The quantum-number of the nuclear angular momentum is a half-integer if the number of particles in the nucleus is odd, and a full-integer (if ascertainable at all) if that number is even.* With the proton-electron scheme, this could not be said.

This difference would give an advantage to the latter scheme, even if the model were no more specific. However, the rules for quantizing orientations of vectors of quantum-number 1/2 require that these shall set themselves either parallel or anti-parallel (in the loose sense) to one another. If I be built up out of such vectors, then necessarily an even number thereof implies a full-integer value and an odd number a half-integer value, and *vice versa*.

This argument for the proton-neutron scheme is therefore strong, though perhaps not so strong as it would be, were not the basis for the test narrowed down by one of the curious empirical rules of the world of atoms: it is found that mostly $(2A - Z)$ is even when A is even, and odd when A is odd. Fortunately there are some exceptions; the famous one is N^{14}, the chief isotope of nitrogen, for which it is certain (from alternating intensities in the band-spectrum) that I is full-integer (unity), whereas $(2A - Z)$ is odd but A is even. This is the only case of its kind, but there are something like ten in which $(2A - Z)$ is even but A is odd, and there is a hyperfine-structure believed to correspond to a half-integer value of I; this seems especially well established for two isotopes of tin and two of mercury. Another and very powerful argument from alternating intensities in band-spectra, unfortunately too long to be expounded here, supports the belief that the nucleus of N^{14} has an even and not an odd number of constituent particles. On the whole it is pretty likely that any nucleus-model providing an even number of particles for an even mass-number and an odd for an odd will always be preferred to any model not having this feature.

The field is now open for interpreting the observed values of I by compounding proton-spins and neutron-spins (or proton-spins and electron-spins), and trying to find reasons for the resultants which are observed. It seems natural enough to have unity for the deuteron (proton and neutron spins parallel), zero for He^4 and C^{12} and O^{16} (spins cancelling each other two by two); but farther along the list

of atoms there are curiosities enough to keep theorists busy, one guesses, for years. I leave this for the future, and close by speaking briefly of the magnetic moments, which by themselves suffice to show that the task will not be easy.

For the magnetic moments attributable to extra-nuclear electrons (or rather to their projections upon the direction of the angular momentum corresponding), we found by theory and by experiment values which we wrote as $g(e/2mc)\ \sqrt{n(n+1)}\ h/2\pi$; here m stands for the mass of the electron; g is 1 for the orbital motion of a single electron, is 2 for the spin of a single electron, and has calculable values not greater than a very few units for the extra-nuclear electron-system as a whole; n is the quantum-number of the associated angular momentum, and is 1/2 for the spin of a single electron and has various values for the other cases.

If now a proton be a simple particle with a spin 1/2, one would expect for its magnetic moment a value differing from that of the electron-spin only by the substitution of m_p the proton-mass for m, therefore about 1840 times smaller. The actual value (page 313) is about three times as great as the expected one, so that the analogy with the electron is good enough to predict the order of magnitude, but is not perfect. This is as surprising a discovery as any that has cropped up in the last several years in the field of atomic physics, and shows that even the supposedly elementary particles still have mysteries for us. The value for the deuteron is a good deal smaller, which with the proton-neutron scheme implies that the neutron has a magnetic moment pointing in the opposite direction to that of the proton. The spins of proton and neutron must, however, be parallel, else their resultant could not be unity, which is the value of I for the deuteron. This brings up a new point: I have not yet said whether magnetic moment and angular momentum should be visualized as parallel or anti-parallel, or better, whether the component of the former along the direction of the latter should be taken as positive or negative. The qualities of the deuteron indicate that whichever is the case with the proton, the opposite is the case with the neutron. Actually it is often possible to tell, from features of hyperfine-structure of which I have said little (cf. footnote 5, p. 302), which of the cases is realized for the nucleus as a whole. For the proton and the deuteron these features are unhappily either absent or inaccessible, and the question remains open. For most of the others which have been analyzed, the magnetic moment is said to be positive, meaning that it is of the sign which one expects, considering the nucleus as a whirling positive charge. For a few nuclei, however, it appears to be negative

Table of Nuclear Spins

This tabulation is presented *sous toutes réserves*. The values are of very different degrees of reliability, but it would be foolhardy for a non-specialist to grade them. They are collected from many sources, including tables by H. E. White (*Introduction to Atomic Spectra*, McGraw-Hill, 1934), by H. Schüler (*ZS. f. Physik*, **88**, 323–335, 1934), by G. Beck (unpublished) and by R. F. Bacher (unpublished), and a number of papers which have appeared in the last fourteen months in *Die Naturwissenschaften*, the *Physical Review*, and other journals.

Values obtained by magnetic-deflection methods are marked R; those deduced from alternating intensities in band-spectra are labelled B; those marked H or not marked at all are inferred from optical observations on hyperfine-structure. Two values for a single mass-number signify either data compatible with both, or discordance between two authorities. X signifies that hyperfine-structure has been sought but not found, which may mean that the spin is zero, but on the other hand may mean that the magnetic moment is too small to produce an observable spread of the hyperfine pattern. When a value of zero for spin is deduced from band-spectra, it is not subject to that doubt.

A	Z	Element	I	A	Z	Element	I
1	1	H	1/2 B	119	50	Sn	1/2
2	1	H	1 B	121	51	Sb	5/2
4	2	He	0 B	123	51	Sb	5/2, 7/2
6	3	Li	X	127	53	I	5/2, 9/2
7	3	Li	3/2 B, H, R	129	54	Xe	1/2
9	4	Be	X	131	54	Xe	3/2
12	6	C	0 B	133	55	Cs	7/2 R, H
14	7	N	1 B	136	56	Ba	X
16	8	O	0 B	137	56	Ba	5/2
19	9	F	1/2	138	56	Ba	X
20	10	Ne	X	139	57	La	5/2, 7/2
22	10	Ne	X	141	59	Pr	5/2
23	11	Na	3/2 R, H	144	62	Sa	X
27	13	Al	1/2	147	62	Sa	$\neq$ 0
31	15	P	1/2 B	148	62	Sa	X
32	16	S	X	149	62	Sa	$\neq$ 0
35	17	Cl	5/2 B	150	62	Sa	X
39	19	K	3/2 B, R	151	63	Eu	$\neq$ 0
41	19	K	$>$ 1/2 R	152	62	Sa	$\neq$ 0
40	20	Ca	X	153	63	Eu	$\neq$ 0
45	21	Sc	7/2	159	65	Tb	3/2, $\geq$ 5/2
51	23	V	$\geq$ 5/2	165	67	Ho	7/2, $\geq$ 5/2
55	25	Mn	5/2	169	69	Tu	1/2
59	27	Co	7/2	175	71	Lu	5/2, $\geq$ 3/2
63	29	Cu	3/2	177	72	Hf	1/2, 3/2
65	29	Cu	3/2	178	72	Hf	X
67	30	Zn	3/2	179	72	Hf	1/2, 3/2
69	31	Ga	3/2	180	72	Hf	X
71	31	Ga	3/2	181	73	Ta	7/2
75	33	As	3/2	185	75	Re	5/2
79	35	Br	3/2, 5/2 B	187	75	Re	5/2
81	35	Br	3/2, 5/2 B	197	79	Au	3/2
83	36	Kr	7/2, 9/2	198	80	Hg	X
85	37	Rb	5/2	199	80	Hg	1/2
87	37	Rb	3/2	200	80	Hg	X
87	38	Sr	3/2?	201	80	Hg	3/2
89	39	Y	1/2	202	80	Hg	X
93	41	Cb	9/2	203	81	Tl	1/2
103	45	Rh	X	204	80	Hg	X
110	47	Cd	X	204	82	Pb	X
111	47	Cd	1/2	205	81	Tl	1/2
112	47	Cd	X	206	82	Pb	X
113	47	Cd	1/2	207	82	Pb	1/2
114	47	Cd	X	208	82	Pb	X
116	47	Cd	X	209	83	Bi	9/2
115	49	In	9/2	231	91	Pa	3/2
117	50	Sn	1/2				

Table of Nuclear Magnetic Moments

The values are expressed in terms of the unit $(eh/4\pi m_p c)$ customary in such tables (though for no very good reason) and sometimes called the "nuclear magneton." They are collected from the same sources as those in the Table of Nuclear Spins, and reduced to at most two significant figures; in cases of discrepancy, Schüler's value is the one usually taken. Those marked S are the ones obtained by Stern and his school; those marked R are obtained by Rabi and his school; those marked H or not at all are deduced from optical observations on hyperfine structure. I have heard the uncertainty of these last put at twenty per cent by one expert in the field, but some authorities would set it even higher. For the isotopes of hydrogen, the sign is unknown.

A	Z	Element		A	Z	Element	
1	1	H	2.5 S	93	41	Cb	3.5
			3.2 R	111	48	Cd	− 0.5
2	1	H	0.5–1.0 S	113	48	Cd	− 0.5
			0.77 R	115	49	In	5.3
7	3	Li	3.3 H, R	117	50	Sn	− 0.9
19	9	F	2.4	119	50	Sn	− 0.9
23	11	Na	2.1	121	51	Sb	2.7
27	13	Al	1.9	123	51	Sb	2.1
39	19	K	0.4 H, R	129	54	Xe	− 1
45	21	Sc	3.5	133	55	Cs	2.5
59	27	Co	2.8	137	56	Ba	1.0
63	29	Cu	2.7	197	79	Au	0.2
65	29	Cu	2.7	199	80	Hg	0.5
69	31	Ga	2.1	201	80	Hg	− 0.6
71	31	Ga	2.7	203	81	Tl	1.5
75	33	As	0.9	205	81	Tl	1.5
83	36	Kr	− 1.0	207	82	Pb	0.5
85	37	Rb	1.5	209	83	Bi	3.6
87	37	Rb	3.1				

The ratio of the magnetic moments of two isotopes of a single element may sometimes be calculated from hyperfine-structure with less uncertainty of theory than the value of either moment separately. For such ratios the following values are available:

$Cu^{63}/Cu^{65} = 1.00$; $Ga^{69}/Ga^{71} = 1.27$; $Rb^{85}/Rb^{87} = 2.04$; $Cd^{111}/Cd^{113} = 1$;
$Sn^{117}/Sn^{119} = 1$; $Sb^{121}/Sb^{123} = 1.36$; $Hg^{199}/Hg^{201} = -1.11$; $Tl^{203}/Tl^{205} = 1.02$.

The experiments of Rabi's school and those of Stern's indicate a value of about *four* for the ratio of the moments of proton and deuteron; this is substantiated by L. and A. Farkas (*Nature*, **135**, 372; 9 March, 1935) who measure the rates of the reactions whereby ortho-hydrogen and ortho-deuterium transform themselves into the para-forms (and reversely) in the presence of oxygen; from measurements made at three different temperatures they get for the ratio the three values 3.85, 4.03, 4.07, which agree (they say) "within the limits of the experimental error, which is less than five per cent."

(it is perhaps significant that all of these have even atomic numbers and odd mass-numbers).[15] As for its numerical magnitude, I recall that the insufficiencies of the theory render the recorded values—of which there are some twenty-five—subject to much doubt. Such as they are, it is a striking fact that none of them is more than a couple of times as great as that of the proton, and some are a good deal smaller. This injures still more the scheme of constructing nucleus-models out of electrons and protons, for the magnetic moment of a single electron not compensated by some other should of itself make the nuclear moment enormously larger. Yet if we adopt the proton and the neutron as the sole constituents of nuclei for the sake of banishing this difficulty, it will return to haunt us so soon as we attempt to reduce the number of elementary particles by conceiving either the neutron as a proton united with a negative electron, or the proton as a neutron with which a positive electron is combined.

ACKNOWLEDGMENTS

I am much indebted to Professor I. I. Rabi and Dr. R. F. Bacher for having read over the manuscript of this article and having made many helpful comments, and to Professors G. Beck, S. Goudsmit and A. E. Ruark for correspondence.

[15] Two isotopes of Cd (A = 111 and 113), two of Sn (117 and 119), one of Hg (201), one of Kr (83), one of Xe (129).

Ferromagnetic Distortion of a Two-Frequency Wave

By ROBERT M. KALB and WILLIAM R. BENNETT

Frequency components are found for the ferromagnetic induction produced by a small magnetizing force of two incommensurable frequencies. Because of hysteresis the results depend intimately upon the ratios of these frequencies and of their amplitudes. With these ratios as criteria, two solutions are provided, adequate for most modulation problems of this character occurring in the field of communications.

The development is based on Madelung's empirical propositions. From these are deduced the forms of complex hysteresis loops occasioned by two-frequency magnetomotive forces, and from the loops sinusoidal components of the flux wave are derived by means of Fourier's series. The various voltages generated in a coil by such a flux are then calculated and next correlated with analyses for a single applied frequency. The resulting changes in the impedances to the two fundamental frequencies are also evaluated. The most important results are given in graphs and tables.

Experimental data on a number of specimens show close agreement with curves computed by the theory.

The analysis discloses several interesting features. It is shown that Madelung's conclusions imply Rayleigh's law of loop similarity; as a consequence the parameters of a Rayleigh loop suffice to describe a complex loop to the extent that it conforms to Madelung's results. Hysteresis suppression is found not to occur at low fields, although harmonic suppression may. The generated side frequencies of the flux appear in unequal pairs, the lower one being the stronger in each instance. Such inequality is a general property ascribable to the multivaluedness of the loop.

FOR precisely evaluating the performance of communication circuits containing ferromagnetic materials, methods for taking into account the non-linear effects of these materials are needed. To this end there have been devoted certain investigations of the behavior of such materials at the low flux densities usual in communication. These early disclosed that hysteresis is a governing factor for weak fields and led to attempts to solve the problem of its bearing on speech.

The complexity of speech and of hysteretic phenomena has made desirable the use of simple testing methods, which in turn require for their interpretation a quantitative theory. Since tests of this sort are usually made with one or more sinusoidal test waves, a theory of single-frequency magnetic performance has already been evolved as a first step toward fulfilling this need. For many purposes single-frequency tests are inadequate, and two-frequency waves are often used to obtain better information bearing on the design or performance of communication systems. It is the purpose of the present paper to take a further step by furnishing the theory of magnetic behavior under a two-frequency force.

The very simplicity which sometimes makes sinusoidal waves valuable for analyzing or testing the non-linear properties of channels of communication makes such waves worthless when applied in other instances, and more complex test waves must then be employed. The harmonics produced by a sine wave furnish an index of the distorting properties of a system, but the side frequencies produced by two such waves are needed to indicate its modulating properties or to give a measure of the interference between carrier channels. When two waves are used, one may be thought of as the carrier, modulated by the other, whose amplitude is chosen proportionate to the square root of the energy in the more complex modulating wave it represents, or both may be thought of as component waves in the same channel, or as carriers in different channels, producing interference in certain other channels. The amplitudes of the several product frequencies then give a measure of the energy falling in their respective regions of the spectrum under actual operation. The effect of the presence of one fundamental upon the transmission of the other can also be ascertained. Increasing the complexity of the test wave by the superposition of additional frequencies can be seen to afford little added advantage at the cost of much complication, unless the character of the waves actually transmitted is simulated, in which case statistical methods of study can perhaps be applied.

From the foregoing circumstances the utility of information pertaining to the application of two-frequency inputs as well as single-frequency inputs to non-linear circuit elements is apparent; many investigations have been conducted in this field to provide such information. When the current-voltage relation is not single-valued a more intricate treatment is necessary in carrying out the analysis. A general method of attack for double-valued characteristics has been provided and applied to hysteresis loops by E. Peterson [1] to determine the flux in ferromagnetic materials under single-frequency magnetizing forces. The fundamental dependence of loop form upon wave shape precludes immediate extension of Peterson's results to the case of a multi-frequency force except for certain harmonic combinations, one of which he considers.[2] A study of flutter effect has been published by Walter Deutschmann,[3] who analyzed a complex loop made up of straight lines. Both instances serve to emphasize the desirability of a broader investigation of the theoretical aspects of two-frequency magnetization including hysteresis. While no general and rigorous

[1] *B. S. T. J.*, Vol. 7, pp. 762–796, Oct. 1928.
[2] *Ibid.*, p. 773.
[3] *Wiss. Ver.* a.d. Siemens-Konzern, Vol. 8, No. 2, pp. 22–44, 1929; *E. N. T.*, Vol. 6, pp. 80–86, Feb. 1929.

method has yet been developed for handling this problem in a manner analogous to that applicable for a single frequency, practically important cases are solved here by means extensible to more complicated ones.

Considerations of Wave Shape and Loop Forms

Scope of the Two-Frequency Analysis

The envelope of a wave affords a means for classification. Waves whose envelopes change gradually during any oscillation form a class apart from those waves which have envelopes subject to abrupt changes. Members of each class can be segregated into those with envelopes nearly symmetrical with respect to the average magnetizing force, those with envelopes of almost uniform width, and so on. One of the two last-mentioned properties in a wave with a gradually changing envelope is essential to successful analysis by the methods about to be detailed.

The cases of magnetization analyzed, which include all the two-frequency wave shapes that can qualify under the foregoing criterion of tractability to analysis, are the following:

Case 1. The ratio between the geometric and arithmetic means of the amplitudes much smaller than the ratio between the sum and difference of the frequencies.

Case 2. One fundamental frequency high relative to the other; the product of the higher frequency with its amplitude large relative to the product of the lower frequency with its amplitude.

Between the two cases there exist intermediate ratios of frequencies and amplitudes over which the theory does not extend; however, the most frequent problems are usually entirely within the domain of a single case. The inequalities involved in the foregoing case definitions are not susceptible to simple explicit statement as limiting numerical ratios, in advance of the development of the theory. Much depends upon the accuracy required in predicting performance. From data supplied in the paper, following the theory, it is possible to determine the practical limitations of the mathematical treatment.

Formation of Complex Loops

Considering the complex hysteresis loops arising from multi-frequency magnetizing forces to be many-valued characteristics for determining the flux density, it is pertinent to study their formation and to correlate their parameters in so far as may be possible with those of single-frequency loops. The latter at low fields are known

to consist approximately of two parabolic branches, the exact shape of each dependent upon its point of origin. For fields confined somewhat below maximum permeability, the situation customarily obtaining in communication circuits, this representation has proved to be sufficiently exact to warrant the neglect of higher order terms in analyses. In conformity, terms of higher than second order will not be retained in equations used here; the development so presented can be extended to include them without other change in procedure. The induction at any point on a simple loop centered on the origin is expressed as a function of the instantaneous magnetizing force h by means of a formula developed by Peterson,

$$B = (a_{10} + a_{11}H)h \pm a_{02}(H^2 - h^2). \tag{1}$$

The upper sign of the double sign is used for the descending (upper) branch of the loop, and the lower sign for the ascending (lower) branch. H is the maximum magnetizing force and the coefficients are constants of the ferromagnetic material, determinable by single-frequency measurements. They have the following significance: a_{10} is the initial permeability, a_{11} the rate of change of permeability with magnetizing force, and a_{02} a factor of proportionality between the hysteresis loss and the cube of the maximum magnetizing force. The concepts in terms of which these parameters are defined acquire extended meanings for complex loops.

In the absence of an adequate theory of ferromagnetism the question of whether branches of complex loops and of simple loops have similar forms must be answered by experiment. The steady state of retracing alternately the two branches of a simple loop may eventuate in a different relation of B versus h than results from the first cycle; such a condition would mean that transient branches compose the complex loop, inasmuch as it is not retraced. It is also possible that the biasing effect of one sinusoidal component of the magnetizing force upon the other might cause the branches of the two types of loops to be dissimilar. The coefficients which specify the branches of the simple loop are evaluated with it centered at the origin of the B–h plane, using single-frequency methods, and cannot be assumed *a priori* to apply to to other situations, or to an unrepeated branch.

According to experiments by R. Goldschmidt [4] the superposed field necessary to cause much change of either the shapes or axial slopes of loops exceeds the weak fields to which this development is limited. Likewise, Lord Rayleigh [5] in his original investigation found small super-

[4] *Zeits. f. Techn. Physik*, Vol. 11, pp. 8–12, 1930.
[5] *Phil. Mag.*, Vol. 23, 1887.

posed fields to have no measurable effect upon the form of the loops for the specimens he investigated. Based upon these results, the branches of simple loops can be taken to be independent of their location in the B–h plane, and in so far as the complex magnetizing force gives rise to branches of simple loops, they have the same shape they would have in such a loop centered at the origin.

The formation of complex loops has been determined experimentally by E. Madelung.[6] He found that after reversing the magnetizing force at any point along the branch of a hysteresis loop a new curve is traced which, if continued, passes through the tip of the loop. A second reversal before the tip is reached causes another new curve to be traced back to the point of reversal on the original branch, which is followed thenceforth as if the two reversals have not occurred. The return to a reversal point makes all subsequent traces of the loop the same as if no changes of magnetizing force intervened between the two transits through that point. Any branch of the complex loop is then, in accordance with Madelung's determinations, completely specified by two points of reversal—the one from which it starts and the one through which it must pass if continued far enough; after passing the latter point it becomes the continuation of another branch similarly specified by different points of reversal.

The foregoing principles furnish sufficient information for deducing the form of the branches of complex loops. If such a branch be extended to one of the reversal points defining it and a trace then be carried back to the other, the loop so formed will be retraced by repeating the cycle. As these repetitions can be carried on indefinitely, the path must comprise a simple loop, the branch of the complex loop forming a portion of it. Every complex loop can therefore be considered as composed of adjoined sections of simple loops. Each branch of the complex loop is representable on suitably transformed axes by formula (1) with H taken as half the change in magnetizing force between the reversal points which specify the branch. In general, a different pair of axes will be required for each branch; they can subsequently be referred to a common origin.

The application of this analysis to the complex loop discloses the requirement that the relation $a_{11} = 2a_{02}$ must be true if Madelung's propositions are to hold, because the values of H differ for the two branches of a subsidiary loop. If this equality is not satisfied, the return to the original branch does not take place at the point of departure. Madelung's observations do not include this possibility,

[6] *Annalen der Physik*, Vol. 17, pp. 861–890, 1905. See also *Handbuch der Physik*, Vol. 15, pp. 106–107, Berlin, 1927.

and sufficient evidence of behavior for extending them in such circumstances is not available at present. Some experiments by Lehde[7] and others indicate that subsidiary loops do not quite close at their junctions with the major loop and often show departures both ways on different parts of the same complex loop. This sort of behavior is not explainable by inequality between a_{11} and $2a_{02}$ for the subsidiary loops, as it would cause the departure of one specimen to be always the same way depending upon which quantity was the larger. In all cases, even near saturation, Lehde's results show this departure to be small and the connecting branches between successive subsidiary loops to form approximately a simple loop. On this basis Madelung's results can be considered confirmed to a sufficient degree of approximation.

The ratio a_{11}/a_{02}, which has been taken as a measure of the validity of Rayleigh's relation in single-frequency theory, becomes a criterion of the usefulness of Madelung's propositions concerning loop form in multi-frequency theory. Those substances which most closely accord with Rayleigh's analysis can also be expected to be in best agreement with Madelung's results. The relation between coefficients required on the basis of Madelung's and Rayleigh's experiments will be used hereafter to simplify the analysis. The simplification will be evidenced by the customary nomenclature, in which

$$\mu_0 \equiv a_{10}, \quad \nu \equiv a_{02} = \tfrac{1}{2}a_{11}.$$

Any attempt to distinguish here between the two latter constants would be meaningless because beyond the scope of Madelung's empirical rules. Fortunately the constants of most commercial materials conform closely to the above equality.

Types of Two-Frequency Loops

The aspect of a hysteresis loop formed by a two-frequency wave changes greatly with the frequencies and their amplitudes. Different pairs of frequencies having equal ratios give rise to families of loops which are identical except as affected by eddy currents. These, for the purpose of this study, are supposed to be so small that the flux is substantially uniform over a cross-section of the magnetic circuit. If the two frequencies have a common source or are synchronized, the hysteretic phenomena are singly periodic and subject to simpler treatment than developed here for independent sources.

For detailed analysis of the effects of hysteresis with two applied frequencies from independent sources, the phase angles of both may

[7] *Rev. of Sci. Instr.*, Vol. 2, pp. 16–43, Jan. 1931.

be taken as zero, equivalent to measuring the time from a point where the two components of the magnetizing force become maximum simultaneously. Then

$$h = P \cos pt + Q \cos qt \tag{2}$$

is the instantaneous magnetizing force. Phase angles can be made arbitrary by replacing pt and qt by $pt + \theta_p$ and $qt + \theta_q$, respectively, in this equation and the subsequent results.

The configurations of complex loops can be followed by altering single-frequency loops to accord with the results of investigations of loop formation. If to a single-frequency magnetizing force a relatively small one of slightly lower frequency is added, the resultant is an oscillatory force whose peaks undulate around the value reached by those of the original wave. Their maximum will be the sum of the amplitudes of the two components and their minimum the difference. On a hysteresis loop (Fig. 1*a*) this means that portions between successive reversal points will differ slightly from one another, being formed approximately as if belonging to successively smaller loops until a minimum peak is passed, thenceforth as if belonging to successively larger loops, and so on cyclically. Such behavior is sketched in the figure.

As the amplitude of the lower frequency component of the magnetizing force is increased the undulations become more pronounced and the preceding picture more inexact. When both amplitudes are equal the envelope of the resultant magnetizing force vanishes periodically and the portion of the hysteresis loop formed while this envelope goes from its maximum to zero is in the nature of a spiral (Fig. 1*b*), a similar curve being developed outwardly as the envelope increases again to its maximum. Provided only that successive peaks of the magnetizing force do not differ greatly in magnitude, each portion of such a loop between adjacent reversal points may be assumed to have the form of a branch of a single-frequency hysteresis loop having its point of origin coincident with that of the portion of the complex loop. Then the induction may be derived from the magnetizing force by the use of single-frequency data in accordance with the known manner of formation of complex loops.

If now the amplitude of the higher frequency component be decreased to a relatively small value, the undulations in the envelope subside, and a condition similar to the original one is seen to obtain. This time, however, the amplitude of the lower frequency will be found to be the one about which these undulations occur, and the characteristic will again look like that in Fig. 1*a*.

As long as one frequency is less than twice the other the undulations in the peaks of the magnetizing force will be regular and gradual. If the higher frequency be raised to more than twice the lower, the undulations become more abrupt and complex, and increasingly so

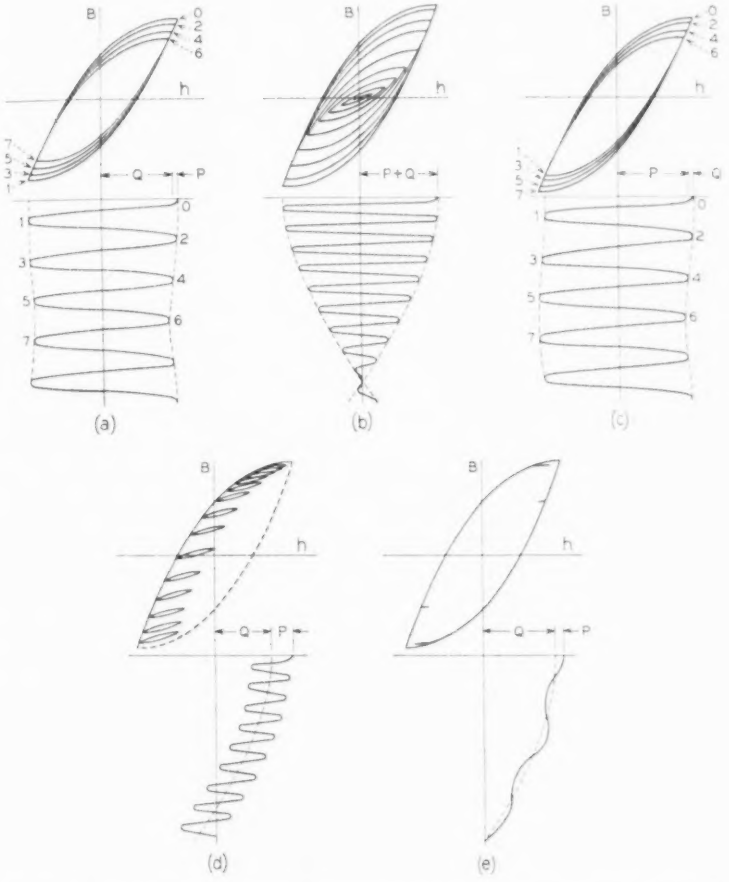


Fig. 1—Types of hysteresis loops characteristic of a two-frequency magnetizing force.

as it is raised still more. The complexity attendant upon the formation of the hysteresis loops becomes greater and the simplifying artifices heretofore suggested no longer apply.

When one frequency has become several times the other, the higher

frequency component must have one of its maxima very near each maximum of the lower one so that a maximum value practically equal to the sum of the amplitudes of the two components is reached every cycle of the lower one; likewise for the minima. Between these two extremes other reversals in the magnetizing force would be expected to cause the hysteresis loop to comprise additional small loops not necessarily closed. Experiments confirm this conjecture and indicate that not only for weak fields but even for fields near saturation the small loops nearly close and the paths between them are traced nearly the same as portions of a large loop.

If the amplitude of the higher frequency component is considerably the larger, all the loops composing the characteristic are virtually the same size and shifted slightly with respect to each other on account of the lower frequency component. The characteristic will be that depicted in Fig. 1*c*.

When the amplitudes of the two components are not grossly unequal, the hysteresis loop is of the type represented by Fig. 1*d*. Small loops formed when the magnetizing force is near an extreme value are longer than those formed when it is near zero, aside from any effects of superposition, because of its different rates of change in the two positions. In general these loops will not occur in the same place for different cycles and the distances between them will be lessened by increasing the higher frequency. By considering the complex loop to consist of a major loop, such as a single frequency would generate, encompassing a number of minor loops, the induction may be derived from the magnetizing force since the form of the minor loops is known. When these are not too widely spaced, each may be assigned a mean position in a loop fixed for all time and the induction calculated therefrom. It is evident that such an undertaking is vastly more complicated than the ones suggested heretofore and that unlike them it requires information in addition to that obtained from single-frequency measurements performed with no superposed field.

By decreasing the amplitude of the higher frequency component of the magnetizing force, the amplitudes of the minor loops may be reduced until those in the neighborhood of zero magnetizing force vanish entirely. Further decrease of the same component causes more and more of the minor loops to disappear, so that finally only a few small ones remain at each end of the major loop. This condition is shown in Fig. 1*e*. As the wave form of the induction is only slightly affected by the presence of these loops, they may safely be omitted and the characteristic simply taken as the major loop, determined from single-frequency results alone.

The point where the minor loops first vanish and the requirement that they vanish at all may be readily determined. Two adjacent extremes of the magnetizing force approach a common value as the amplitude of the higher frequency component is decreased. When they attain this value, the slope is no longer reversed between them and consequently no minor loop is formed. The instantaneous magnetizing force is expressed by equation (2), and in this instance $p \gg q$. The minor loops first appear where $dh/dt = 0$ when $pt = -\frac{\pi}{2}$. Solved simultaneously, these equations yield

$$qt = \sin^{-1} \frac{Pp}{Qq}, \tag{3}$$

where the minor loops vanish. They reappear at an equal negative value of $\sin qt$, and other intervals during which they vanish are apparent from symmetry. If $(Pp/Qq) > 1$, no real solution of equation (3) exists, so the minor loops do not vanish anywhere. The appearance and non-appearance of minor loops is seen to be governed by the ratio of amplitudes in the same way as by the ratio of frequencies as long as the restriction on the latter is observed, and the product of these ratios

$$\kappa = \frac{Qq}{Pp}$$

determines the type of hysteresis loop (1*d* or 1*e*, or an intermediate form) obtained.

Induction with a Two-Frequency Magnetizing Force

General Expression for the Induction

As a function of time, the induction for any type of loop described will consist of intermodulation products of the two fundamental frequencies. Because of the kind of symmetry the characteristic has these products will include only the odd orders, but because it is a loop, quadrature terms must be expected. The induction then will ultimately be in the form

$$B = \sum_{m=0}^{\infty} \sum_{n=-\infty}^{\infty} [c_{mn} \sin (mp + nq)t + d_{mn} \cos (mp + nq)t] \tag{4}$$

with all even order coefficients zero. The odd order coefficients remain to be determined from the hysteresis loop. In doing this only third order products in addition to the fundamentals will be evaluated explicitly, as they are stronger than the higher orders and therefore

of most importance in measurements of distortion. Any higher order products would be less precisely evaluated, more numerous, and probably of less interest; their computation is clearly evident as an extension of the processes later carried out.

The Multi-Branched Hysteresis Loop

Under certain circumstances mentioned earlier, the portions of a complex loop between adjacent reversal points are representable by

$$B = (\mu_0 + 2\nu H)h \pm \nu(H^2 - h^2). \tag{5}$$

Referred to the origin of the B–h plane, the induction on such a portion originating at the jth reversal point from $t = 0$ is

$$B_j = G_j + \mu_0(h - H_j) - (-1)^j\nu(h - H_j)^2. \tag{6}$$

Here H_j is the magnetizing force and G_j the induction at the jth reversal point. The latter quantity satisfies the difference equation

$$G_j = G_{j-1} + \mu_0(H_j - H_{j-1}) + (-1)^j\nu(H_j - H_{j-1})^2,$$

arrived at by evaluating the induction on the $(j-1)$st branch at the jth reversal point. Subject to the initial condition

$$G_0 = (\mu_0 + 2\nu H_0)H_0,$$

this can be solved by the method of successive substitutions; the solution is

$$G_j = \mu_0 H_0 + 2\nu H_0^2 + \mu_0 \sum_{i=1}^{j} (H_i - H_{i-1}) + \nu \sum_{i=1}^{j} (-1)^i (H_i - H_{i-1})^2. \tag{7}$$

The foregoing expressions define the induction everywhere on the complex loop. Equations (7) and (6) combined to eliminate G_j give

$$B_j = \mu_0 h + (-1)^j 2\nu H_j h + (-1)^j \nu(H_j^2 - h^2). \tag{8}$$

The problem remaining is to develop this equation into the equivalent of equation (4).

The instantaneous magnetizing force is

$$h = P \cos pt + Q \cos qt. \tag{2}$$

By a trigonometric transformation this may be put in the form

$$h = \sqrt{(P+Q)^2 - 4PQ\sin^2\frac{p-q}{2}t} \times \cos\left[\frac{p+q}{2}t + \tan^{-1}\left(\frac{P-Q}{P+Q}\tan\frac{p-q}{2}t\right)\right], \quad (9)$$

which is sometimes more convenient. The envelope of the wave is represented by the two branches of the radical. If its magnitude does not change much between adjacent maxima and minima of the wave, these extremes lie close to the points of tangency between the wave and its envelope; the latter condition is the one necessary in order that the envelope may be used to evaluate the extreme magnetizing force H_j.

This force acquires its values at the reversal points, which are situated at the zeros of dh/dt. Put into a form like equation (9) by the same transformations as were used above, this derivative is

$$\frac{dh}{dt} = -\sqrt{(Pp+Qq)^2 - 4PpQq\sin^2 st} \times \sin\left[rt + \tan^{-1}\left(\frac{Pp-Qq}{Pp+Qq}\tan st\right)\right], \quad (10)$$

where $2r = p + q$, $2s = p - q$. Except when $\kappa = 1$, this vanishes only at

$$rt + \tan^{-1}\left[\frac{1-\kappa}{1+\kappa}\tan st\right] = j\pi, \quad (11)$$

j integral or null. Substituting equation (11) into equation (9) yields the magnetizing force at the jth reversal from $t = 0$:

$$H_j = \sqrt{(P+Q)^2 - 4PQ\sin^2 st} \times \cos\left[j\pi + \tan^{-1}\left(\frac{1-k}{1+k}\tan st\right) - \tan^{-1}\left(\frac{1-\kappa}{1+\kappa}\tan st\right)\right],$$

letting $k = Q/P$. Upon combining the arc tangents this becomes

$$H_j = P\sqrt{(1+k)^2 - 4k\sin^2 st} \times \left[\cos j\pi - \tan^{-1}\frac{2(k-\kappa)\tan st}{(1+k)(1+\kappa) + (1-k)(1-\kappa)\tan^2 st}\right]. \quad (12)$$

According to equation (12) the magnetizing force and its envelope are tangent at reversal points provided the arc tangent is always zero. This it is if $k = \kappa$, a trivial solution. By certain choices of these two parameters, however, it is possible to keep the angle between any reversal point and the nearest extreme of the magnetizing force from exceeding any prescribed limit. The maximum value of the angle in

equation (12) is

$$2 \tan^{-1} \sqrt{\frac{(1-k)(1+\kappa)}{(1+k)(1-\kappa)}} - \frac{\pi}{2},$$

which can be made small by making k and κ each small compared to unity, or each large compared to unity, or both approximately equal. For the previously excluded instance $\kappa = 1$, this angle can be limited and the last condition fulfilled by keeping k nearly equal to one, as is obvious from the equations. So for each of these three conditions on the parameters, to a definite degree of approximation the envelope at each point of tangency becomes the magnetizing force for the nearest reversal point, a feature useful for the transformation of equation (8) into a function of time.

Calculation of the Induction—Case 1

The three conditions are confluent and will be seen to set the limiting bounds for case 1. When the fundamental frequencies lie close together the ratio of their amplitudes is practically unrestricted; as more widely spaced frequencies are chosen it becomes necessary to require an increased (or diminished) amplitude ratio in order that the phase angle in equation (12) does not exceed the chosen limit. This limit must be such that the cosine of that phase angle is substantially unity.

The maximum magnetizing force is thereupon

$$H_i = (-1)^i P \sqrt{(1+k)^2 - 4k \sin^2 st}. \tag{13}$$

As a periodic even function of st, H_i may be expanded in a Fourier's series

$$H_i = (-1)^i (P+Q) \left[\frac{A_0}{2} + \sum_{\eta=1}^{\infty} A_\eta \cos \eta st \right], \tag{14}$$

where

$$A_\eta = \frac{4}{\pi} \int_0^{\pi/2} \sqrt{1 - k_1^2 \sin^2 \lambda} \cos \eta\lambda d\lambda \tag{15}$$

in terms of the parameter

$$k_1 = \frac{2\sqrt{PQ}}{(P+Q)} = \frac{2\sqrt{k}}{(1+k)} = \frac{2\sqrt{1/k}}{(1+1/k)},$$

which never exceeds unity and which diminishes as k is either increased or decreased from unity. The integral (15) reduces immediately to elliptic form by the substitution $z = \sin \lambda$. All odd order coefficients are zero. For the first three significant values of η:

$$\left.\begin{aligned} A_0 &= \frac{4}{\pi} E_1 \\ A_2 &= \frac{4}{3\pi k_1^2}[(2 - k_1^2)E_1 - 2(1 - k_1^2)K_1] \\ A_4 &= \frac{4}{15\pi k_1^4}[8(2 - k_1^2)(1 - k_1^2)K_1 - (16 - 16k_1^2 + k_1^4)E_1] \end{aligned}\right\} \quad (16)$$

Here K_1 and E_1 are complete elliptic integrals of the first and second kind, respectively, with modulus k_1.

The series (14) may be used to evaluate the variable permeability anywhere on the loop, for upon substitution in equation (8) reference to a particular branch is eliminated by the disappearance of the double sign on the second term.

The square of the maximum magnetizing force needed in the final term of equation (8) comes directly from equation (13); to determine the sign of this term at any instant remains the only problem. Interpretation of $(-1)^j$ seems simple when it is remembered to be positive for decreasing h and negative for increasing h, and therefore an odd function of time. The rate of change of the magnetizing force is

$$\frac{dh}{dt} = -Pp \sin pt - Qq \sin qt,$$

so it follows that

$$\left.\begin{aligned} (-1)^j &= +1, \quad \sin pt + \kappa \sin qt > 0 \\ &= -1, \quad \sin pt + \kappa \sin qt < 0 \end{aligned}\right\}. \quad (17)$$

The solution may be completed by expanding this quantity in a Fourier's series:[8]

$$(-1)^j = \sum_{m=0}^{\infty} \sum_{n=-\infty}^{\infty} A_{mn} \sin (mp + nq)t. \quad (18)$$

When $m = 0$ the summation is to be extended over only positive values of n. With this convention the coefficients are

$$A_{mn} = \frac{1}{2\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} (-1)^j \sin (mx + ny) dx\, dy \quad (19)$$

or, with the use of equation (17),

$$A_{mn} = (-1)^{\frac{m+n-1}{2}} \frac{4}{\pi^2} \int_0^{\pi} dy \int_0^{\cos^{-1}(-\kappa \cos y)} \cos mx \cos ny\, dx, \quad (20)$$

[8] W. R. Bennett, "New Results in the Calculation of Modulation Products," *B. S. T. J.*, Vol. 12, pp. 228–243, April, 1933.

where p and q are to be so assigned that $\kappa \leq 1$. Then the coefficients are all expressible in terms of complete elliptic integrals with modulus κ; these will be designated as K_2 and E_2. Coefficients of even order vanish, while those of the first three odd orders are found to be

$$\left.\begin{aligned}
A_{10} &= \frac{8}{\pi^2} E_2 \\
A_{01} &= \frac{8}{\pi^2 \kappa} [E_2 - (1 - \kappa^2) K_2] \\
A_{\underline{21}} = -A_{21} &= \frac{8}{3\pi^2 \kappa} [(1 - 2\kappa^2) E_2 - (1 - \kappa^2) K_2] \\
A_{\underline{12}} = A_{12} &= \frac{8}{3\pi^2 \kappa^2} [(2 - \kappa^2) E_2 - 2(1 - \kappa^2) K_2] \\
A_{30} &= \frac{8}{9\pi^2} [(7 - 8\kappa^2) E_2 - 4(1 - \kappa^2) K_2] \\
A_{03} &= \frac{8}{9\pi^2 \kappa^3} [(8 - 3\kappa^2)(1 - \kappa^2) K_2 - (8 - 7\kappa^2) E_2] \\
A_{\underline{41}} = -A_{41} &= \frac{8}{15\pi^2 \kappa} [(1 - 16\kappa^2 + 16\kappa^4) E_2 \\
&\qquad - (1 - 8\kappa^2)(1 - \kappa^2) K_2] \\
A_{\underline{14}} = A_{14} &= \frac{8}{15\pi^2 \kappa^4} [8(2 - \kappa^2)(1 - \kappa^2) K_2 \\
&\qquad - (16 - 16\kappa^2 + \kappa^4) E_2] \\
A_{\underline{32}} = A_{32} &= \frac{8}{15\pi^2 \kappa^2} [2(1 + 2\kappa^2)(1 - \kappa^2) K_2 \\
&\qquad - (2 + 3\kappa^2 - 8\kappa^4) E_2] \\
A_{\underline{23}} = -A_{23} &= \frac{8}{15\pi^2 \kappa^3} [(8 + \kappa^2)(1 - \kappa^2) K_2 - (8 - 3\kappa^2 - 2\kappa^4) E_2] \\
A_{50} &= \frac{8}{75\pi^2} [(43 - 168\kappa^2 + 128\kappa^4) E_2 \\
&\qquad - 4(7 - 16\kappa^2)(1 - \kappa^2) K_2] \\
A_{05} &= \frac{8}{75\pi^2 \kappa^5} [(128 - 168\kappa^2 + 43\kappa^4) E_2 \\
&\qquad - (128 - 104\kappa^2 + 15\kappa^4)(1 - \kappa^2) K_2]
\end{aligned}\right\} \quad (21)$$

Negative digits of the subscripts are underscored in the coefficients for lower side frequencies.

Upon putting the various quantities into equation (8) from equations (2), (13), (14), and (18) it thus becomes

$$
\begin{aligned}
B = {} & \mu_0 P \cos pt + \mu_0 Q \cos qt \\
& + \nu P(P+Q) \sum_{m=0}^{\infty} (A_{2m} + kA_{2m+2}) \cos [(m+1)p - mq]t \\
& + \nu P(P+Q) \sum_{m=0}^{\infty} (kA_{2m} + A_{2m+2}) \cos [mp - (m+1)q]t \\
& - \tfrac{1}{4}\nu P^2 \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} (A_{m+2,\,n} - 2A_{mn} + A_{m-2,\,n}) \\
& \qquad \times \sin [mp + nq]t \\
& - \tfrac{1}{4}\nu Q^2 \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} (A_{m,\,n-2} - 2A_{mn} + A_{m,\,n+2}) \\
& \qquad \times \sin [mp + nq]t \\
& + \tfrac{1}{2}\nu PQ \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} (A_{m+1,\,n-1} - A_{m+1,\,n+1} \\
& \qquad + A_{m-1,\,n+1} - A_{m-1,\,n-1}) \sin [mp + nq]t
\end{aligned}
\tag{22}
$$

with the understanding that $A_{rs} \equiv 0$ for $r < 0$ and for $r = 0$, $s < 0$. The first line is the linear portion of the induction, given by a permeability constant at its initial value; the first two summations arise from the variation of the permeability with the maximum magnetizing force; the remaining terms comprise the results of distortion attributable to hysteresis *per se*. The coefficients of the induction, c_{mn} and d_{mn} of equation (4), may now be evaluated by selecting the necessary quantities from equation (22).

General expressions for the coefficients can be evolved in series known as hypergeometric functions. These are all of the type

$$
F(\alpha, \beta; \gamma; z) = 1 + \frac{\alpha\beta}{\gamma}\frac{z}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \cdots : \tag{23}
$$

a particular one is chosen by specifying the parameters. The coefficients needed here are

$$
A_{mn} = \frac{2}{\pi} \frac{\Gamma\left(\frac{m+n}{2}\right) \kappa^n}{\Gamma(n+1)\,\Gamma\left(\frac{m-n}{2}+1\right)} F\left(\frac{m+n}{2}, \frac{n-m}{2}; n+1; \kappa^2\right) \tag{24}
$$

for m and n both positive and $m + n$ odd, and

$$
A_\eta = \frac{\Gamma\left(\frac{\eta+1}{2}\right) k_1^{\eta}}{\Gamma(\eta+1)\,\Gamma\left(\frac{1-\eta}{2}+1\right)} F\left(\frac{\eta+1}{2}, \frac{\eta-1}{2}; \eta+1; k_1^2\right) \tag{25}
$$

for η even. When n is negative, the coefficient may be found by using the relation

$$A_{m\bar{n}} = (-1)^n A_{mn}; \tag{26}$$

for all other values of the subscripts excluded the coefficients are zero. A recurrence formula for computing products of higher order is

$$A_{mn} = \frac{1}{m+n}\left[2\left(\frac{n-1}{\kappa} + (m-1)\kappa\right)A_{m-1,\,n-1} - (m+n-4)A_{m-2,\,n-2}\right] \tag{27}$$

with $m-2$ positive. Comparison of equation (24) with equation (25) reveals that if in the former κ is replaced by k_1,

$$A_\eta = \frac{\pi}{2}A_{1,\eta}, \tag{28}$$

so the equations (21) and the recurrence formula (27) suffice for computing all the coefficients in the series for the induction.

Calculation of the Induction—Case 2

For case 2 the two branches of a minor loop and an adjoined portion of the major loop are combined into a Fourier's series whose coefficients are functions of position in the major loop. By developing these coefficients into Fourier's series, a double series in time is obtained. For this case, as for the other just considered, the induction thus is developed in the form of equation (4) and the coefficients are determined through the third order.

When minor loops are formed throughout the lower frequency cycle, an expression for each minor loop and the portion of the major loop joining it to the next one is found relative to an origin at the junction of the major and minor loops. A succession of such loops is then referred to the origin of the major loop by transforming the coordinates of a typical minor loop, the transformation being a function of the position of the minor loop.

Attention first will be devoted to a single high-frequency cycle occurring while the lower-frequency component of the magnetizing force is decreasing. Let the time of occurrence of the maximum in the higher frequency component of the magnetizing force during this cycle be designated by τ. By restricting application to characteristics with sizeable minor loops, i.e. $\kappa \ll 1$ when $p > q$, consistent with the stipulation that the minor loops do not vanish anywhere, this maximum

can be made practically coincident with the corresponding maximum in the magnetizing force.

Writing

$$t = \tau + \lambda, \tag{29}$$

the lower-frequency component in the vicinity is expressible by the Taylor's series

$$Q \cos qt = Q\left[1 - \frac{q^2\lambda^2}{2!} + \frac{q^4\lambda^4}{4!} - \cdots\right]\cos q\tau - Q\left[q\lambda - \frac{q^3\lambda^3}{3!} + \frac{q^5\lambda^5}{5!} - \cdots\right]\sin q\tau. \tag{30}$$

Over one cycle of the higher frequency this component is very nearly linear, so its variation in this range is

$$-4\Delta = -2\pi\kappa P \sin q\tau. \tag{31}$$

Since τ has been so chosen as to be an integral multiple of $2\pi/p$, when equations (29) and (30) are substituted in the magnetizing force given by equation (2) it becomes

$$h = P \cos p\lambda + Q \cos q\tau - \lambda Qq \sin q\tau. \tag{32}$$

Its value referred to a new set of coordinates, B', h', with their origin at the junction of the major and minor loops is

$$h' = P(1 + \cos p\lambda) - 2\Delta\left(1 + \frac{p\lambda}{\pi}\right). \tag{33}$$

According to Madelung's findings previous minor loops will not influence the one under consideration, so its lower branch will proceed toward the upper tip of the major loop as indicated in Fig. 2. A transformation of equation (1), simplified by the use of Rayleigh's relation, then gives for this branch

$$B_2' = \mu_0 h' + \nu h'^2. \tag{34}$$

The upper branch is

$$B_1' = [\mu_0 + 4\nu(P - \Delta)]h' - \nu h'^2. \tag{35}$$

The small portion of the major loop traversed during the last of the cycle may be expanded in a Taylor's series

$$B' = B'(0) + h'\frac{\partial B_1}{\partial h}\bigg|_{h=h_r} + \frac{h'^2}{2!}\frac{\partial^2 B_1}{\partial h^2}\bigg|_{h=h_r} + \cdots, \tag{36}$$

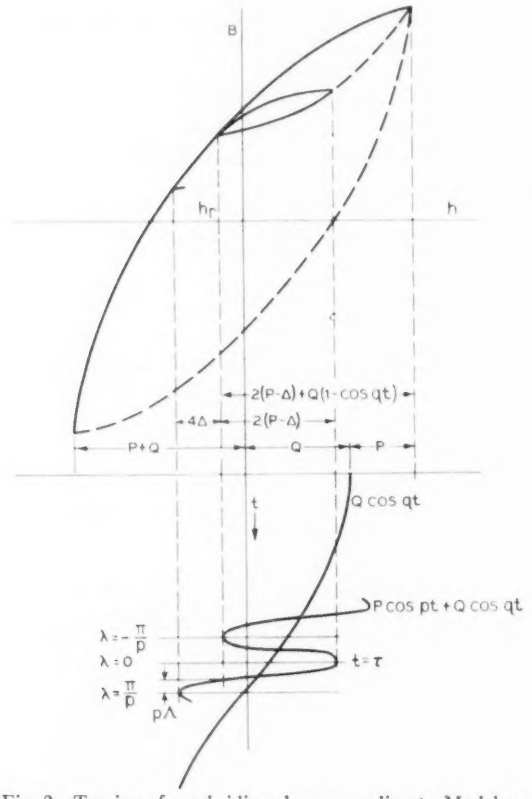


Fig. 2—Tracing of a subsidiary loop according to Madelung.

denoting by h_r the value of h at $h' = 0$, which is

$$h_r = 2\Delta - P + Q \cos q\tau. \tag{37}$$

The upper branch of the major loop as a function of h (not h') is

$$B_1 = (\mu_0 + 2\nu H)h + \nu(H^2 - h^2) \tag{38}$$

with

$$H = P + Q.$$

The expansion (36) is now found to be

$$B' = \{\mu_0 + 2\nu[2(P - \Delta) + Q(1 - \cos q\tau)]\}h' - \nu h'^2. \tag{39}$$

Equations (34), (35), and (39) define the induction in terms of h' over an entire high-frequency cycle; the first is valid for $-\pi < p\lambda < 0$, the second for $0 < p\lambda < \pi - p\Lambda$, and the last for $\pi - p\Lambda < p\lambda < \pi$.

For combining these three expressions a Fourier's series may be developed applicable over the entire interval $-\pi < p\lambda < \pi$, and the induction expressed by this series can be referred to the central axes, B, h, of the major loop by including in its constant term the value of the induction at the junction of the major and minor loops. The series is

$$B_1 = \tfrac{1}{2}b_0' + b_1' \cos p\lambda + b_2' \cos 2p\lambda + b_3' \cos 3p\lambda + a_1' \sin p\lambda + a_2' \sin 2p\lambda + a_3' \sin 3p\lambda. \quad (40)$$

The coefficients are determined by the integrals

$$a_n' = \frac{1}{\pi}\int_{-\pi}^{\pi} B \sin np\lambda \, d(p\lambda), \qquad b_n' = \frac{1}{\pi}\int_{-\pi}^{\pi} B \cos np\lambda \, d(p\lambda), \quad (41)$$

where

$$B = B' + B_r.$$

B_r is the induction at the junction of the major and minor loops, found by inserting equation (37) in equation (38). The resulting quantity together with expressions previously found for B' over various parts of the cycle are substituted in the integrands of equation (41), and the integrations are performed, using h' given by equation (33). All terms of order higher than the third and those containing the square of the frequency ratio as a factor are rejected as they occur. The resulting coefficients are functions of $q\tau$ both explicitly and also through Δ and $p\Lambda$.

To determine $p\Lambda$ as a function of $q\tau$, the vanishing of h' at the tip of the minor loop on the major loop gives an equation for use along the descending branch of the major loop. By equation (33)

$$P(1 - \cos p\Lambda) = 4\Delta\left(1 - \frac{p\Lambda}{2\pi}\right),$$

$$(1 - \cos p\Lambda) = (2\pi - p\Lambda)\kappa \sin q\tau. \quad (42)$$

An approximation to the general solution can be got by transforming equation (42) into a quadratic algebraic equation and solving. This is done by means of the first two terms of the power series expansion for $\cos p\Lambda$. The approximation will be best for small values of the angle, but very good over all its admissible range. This reduction gives

$$(p\Lambda)^2 + 2\kappa \sin q\tau \, (p\Lambda) - 4\pi\kappa \sin q\tau = 0, \quad (43)$$

the roots of which are

$$p\Lambda = -\kappa \sin q\tau \pm \sqrt{\kappa^2 \sin^2 q\tau + 4\pi\kappa \sin q\tau},$$

the positive one being that sought. By expanding the quantity under the radical according to the binomial theorem this root reduces to

$$p\Lambda = 2\sqrt{\pi}\,\sqrt{\kappa \sin q\tau} - \kappa \sin q\tau \tag{44}$$

when higher powers of κ are dropped.

The coefficients of the series then become, using the value of Δ from equation (31) and the value of $p\Lambda$ from equation (44),

$$\begin{aligned}
\frac{1}{2} b_0' &= \mu_0 Q \cos q\tau + 2\nu(2P + Q)Q \cos q\tau + \left(\pi - \frac{4}{\pi}\right) \nu PQ \frac{q}{p} \sin q\tau \\
&\quad + 2\pi\nu Q^2 \frac{q}{p} \sin q\tau\,(1 - \cos q\tau) + \nu Q^2 \sin^2 q\tau, \\
a_1' &= -2\mu_0 Q \frac{q}{p} \sin q\tau - 8\nu PQ \frac{q}{p} \sin q\tau + \frac{8}{3\pi} \nu P^2, \\
b_1' &= \mu_0 P + 2\nu P^2, \\
a_2' &= \mu_0 Q \frac{q}{p} \sin q\tau + \frac{14}{3} \nu PQ \frac{q}{p} \sin q\tau, \\
b_2' &= -\frac{40}{9\pi} \nu PQ \frac{q}{p} \sin q\tau, \\
a_3' &= -\frac{2}{3} \mu_0 Q \frac{q}{p} \sin q\tau - \frac{8}{15\pi} \nu P, \\
b_3' &= 0.
\end{aligned} \tag{45}$$

The relation (29) can be used to return equations (45) to the general time coordinate. Replacing τ by $t - \lambda$ gives

$$\begin{aligned}
\cos q\tau &= \cos q\lambda \cos qt + \sin q\lambda \sin qt, \\
\sin q\tau &= \cos q\lambda \sin qt - \sin q\lambda \cos qt.
\end{aligned}$$

As $|\lambda|$ never exceeds π/p these equations can be simplified to

$$\begin{aligned}
\cos q\tau &= \cos qt + \frac{q}{p}\,(2 \sin p\lambda - \sin 2p\lambda) \sin qt, \\
\sin q\tau &= \sin qt - \frac{q}{p}\,(2 \sin p\lambda - \sin 2p\lambda) \cos qt,
\end{aligned} \tag{46}$$

using the first terms of the Fourier's series for $\sin q\lambda$ in multiples of $p\lambda$. Upon substitution of equation (29) trigonometric functions of $p\lambda$ become the same functions of $(pt - 2j\pi)$, j an integer, since τ is defined as an even integral multiple of $2\pi/p$. The phase angle is

therefore immaterial, so $p\lambda$ can be replaced by pt in equations (40) and (46). Combination of equations (45) and (46) with equation (40) results in an expression of the induction on the upper half of the complex loop in terms of time.

For the lower half of the loop a mean position must be found. Having started with incommensurable frequencies at zero phase angles, the reversals at the lower tip of the major loop will occur at

$$pt = 2m\pi + R,$$

where $0 < R < 2\pi$ and all values of R within these limits are equally probable. The lower frequency component at the instant of reversal will have a time angle

$$qt = (2n + 1)\pi + S,$$

where $-(q/p)\pi < S < (q/p)\pi$ and all values of S within the limits are equally probable. The expected medians of the time angles are therefore $(2m + 1)\pi$ and $(2n + 1)\pi$ for the higher and lower frequency components respectively. These values and the point symmetry of the characteristic specify that the induction during the ascendency of the lower-frequency component of the magnetizing force will be equal in magnitude and opposite in sign to the induction for the descent with the phase of each component increased by its expected median. The induction for an increasing lower-frequency component is therefore given by

$$B_2[pt, qt] = -B_1[pt + \pi, qt + \pi], \tag{47}$$

where the right-hand member is evaluated for a decreasing lower-frequency component.

The coefficients of equation (40) may be altered accordingly to furnish a set for use when the lower-frequency component is increasing by replacing qt by $(qt + \pi)$ and pt by $(pt + \pi)$. In series form, then, the induction on the lower half of the loop is

$$B_2 = \tfrac{1}{2}b_0'' + b_1'' \cos pt + b_2'' \cos 2pt + b_3'' \cos 3pt + a_1'' \sin pt + a_2'' \sin 2pt + a_3'' \sin 3pt. \tag{48}$$

One pair of coefficients is necessary to specify completely the amplitude of each component of the induction when it is split into in-phase and quadrature terms harmonic in pt. Coefficients of corresponding terms in equations (40) and (48) are all functions of qt, each series applying over one half of a lower-frequency cycle. Each pair of coefficients can therefore be developed into a Fourier's series in qt, so that the single expression

$$B = \tfrac{1}{2}b_0 + b_1 \cos pt + b_2 \cos 2pt + b_3 \cos 3pt + a_1 \sin pt + a_2 \sin 2pt + a_3 \sin 3pt \quad (49)$$

defines the induction everywhere.

The coefficients of equation (49) are given by the expressions

$$\begin{aligned} b_n &= \frac{1}{2}[b_n' + b_n''] + \frac{2}{\pi}[b_n' - b_n''] \sum_{m=0}^{\infty} \frac{\sin (2m+1)q\tau}{(2m+1)}, \\ a_n &= \frac{1}{2}[a_n' + a_n''] + \frac{2}{\pi}[a_n' - a_n''] \sum_{m=0}^{\infty} \frac{\sin (2m+1)q\tau}{(2m+1)} \end{aligned} \quad (50)$$

After putting the values of the primed coefficients into the respective terms, changing the arguments of functions of $q\tau$ to qt by using equations (46), and expanding the result into multiple angles, there remains when terms beyond the third order are dropped

$$\begin{aligned} B = {} & [\mu_0 + 2\nu P]P \cos pt + \frac{8}{3\pi}\nu[P + 3\kappa Q]P \sin pt \\ & + \left[\mu_0 + 2\nu \left(2 + k - \frac{4}{3}\kappa \right) P \right] Q \cos qt \\ & + \frac{8}{3\pi}\nu \left[k + \frac{3\pi^2}{4}\kappa + \frac{3}{2}\left(\frac{\pi^2}{4} - 1 \right)\frac{\kappa}{k} \right] PQ \sin qt \\ & - \frac{1}{3}\nu[1 - 6k]\kappa P^2 \cos (2p + q)t \\ & - \frac{40}{9\pi}\nu \left[1 + \frac{3}{10}k \right] \kappa P^2 \sin (2p + q)t \\ & + \frac{1}{3}\nu[1 - 6k]\kappa P^2 \cos (2p - q)t \\ & + \frac{40}{9\pi}\nu \left[1 - \frac{3}{10}k \right] \kappa P^2 \sin (2p - q)t \\ & + \frac{32}{5\pi^2}\nu\kappa PQ \cos (p + 2q)t \\ & - \frac{2}{3\pi}[\mu_0 + 2\nu(2 + 3k)P]\kappa P \sin (p + 2q)t \\ & - \frac{32}{5\pi^2}\nu\kappa PQ \cos (p - 2q)t \\ & - \frac{2}{3\pi}[\mu_0 + 2\nu(2 + 3k)P]\kappa P \sin (p - 2q)t \\ & - \frac{4}{3\pi}\left[\mu_0\kappa - \frac{2}{5}\nu P \right] P \sin 3pt \\ & + \frac{8}{5}\nu\kappa PQ \cos 3qt - \frac{8}{15\pi}\nu Q^2 \sin 3qt. \end{aligned} \quad (51)$$

Recapitulation of Principal Results

General formulæ have now been made available for calculating the flux density over a wide range of conditions of two-frequency magnetization. For many ordinary purposes a table or graph of some of the results is convenient; useful ones are therefore included.

The hypergeometric expansions in the coefficients of case 1 can be put to further use to examine the behavior of the induction for special ratios of fundamental amplitudes and frequencies. When $k \ll 1$ and $\kappa \ll 1$ the coefficients of the several frequencies in the induction reduce to simple, rational, algebraic expressions of the amplitudes. These coefficients likewise reduce when $k \doteq 1$ and $\kappa \doteq 1$, since

$$F(\alpha, \beta; \gamma; 1) = \frac{\Gamma(\gamma)\Gamma(\gamma - \beta - \alpha)}{\Gamma(\gamma - \alpha)\Gamma(\gamma - \beta)}.$$

Third order coefficients for case 1 with restricted parameters, and also those for case 2 are tabulated in the accompanying table through terms in the lowest power of the smaller amplitude. Underlined subscripts distinguish lower side frequencies; a bar under a digit indicates it is to be taken with a negative sign.

When the ratio of amplitudes is unrestricted, graphs of the coefficients which specify the induction enable their magnitudes to be determined most readily. The strongest products are found to be the third order lower side frequencies; Fig. 3, calculated by A. G. Tynan, may be used to get both components of either of these. The corresponding upper side frequencies are almost as strong; their amplitudes can be found from the figure by virtue of the relations $c_{12} = c_{\underline{12}}$ and $c_{21} = c_{\underline{21}}$, since their other components are zero.

By interchanging P and Q and likewise p and q in either the table or the graphs, the subscripts are also reversed. In the table the inequalities restricting the columns are reversed thereby, since the interchanged quantities are arbitrarily assignable. This procedure does not extend the applicability of either case, but does permit the application of case 1 directly to both extremes of the amplitude ratio and the use of the curves to evaluate the amplitudes of both lower side frequencies. The scope of the table and curves given has been extended in this manner. A field of usefulness sufficiently extensive for most purposes of present-day communication is thereby achieved.

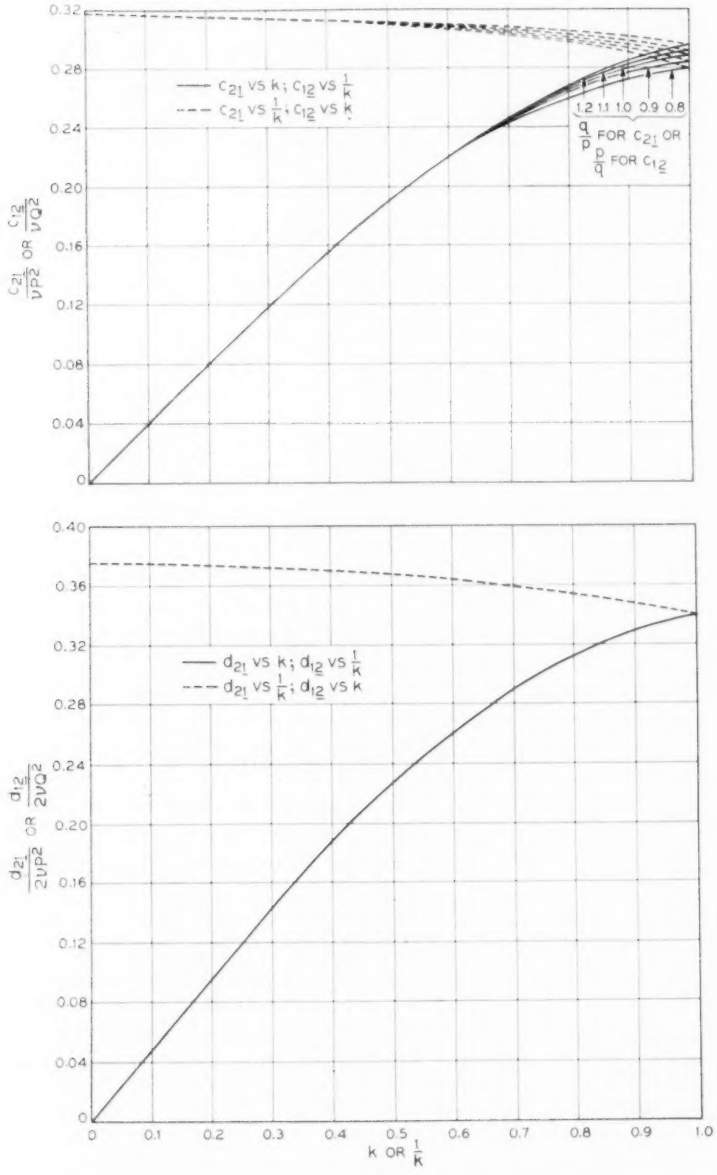


Fig. 3—Components of third order lower side frequencies—case 1.

TABLE I

	Case 1a $\kappa \ll 1$ $k \ll 1$	Case 1b $\kappa \doteq 1$ $k \doteq 1$	Case 1c $\kappa \gg 1$ $k \gg 1$	Case 2 $\kappa \ll 1$ $k \gg \kappa$
c_{10}	$\frac{8}{3\pi}\nu P^2$	$\frac{128}{9\pi^2}\nu PQ$	$\frac{4}{\pi}\nu PQ$	$\frac{8}{3\pi}\nu P^2$
d_{10}	$[\mu_0 + 2\nu P]P$	$\left[\mu_0 + \frac{32}{3\pi}\nu P\right]P$	$[\mu_0 + 3\nu Q]P$	$[\mu_0 + 2\nu P]P$
c_{01}	$\frac{4}{\pi}\nu PQ$	$\frac{128}{9\pi^2}\nu PQ$	$\frac{8}{3\pi}\nu Q^2$	$\frac{8}{3\pi}\nu Q^2$
d_{01}	$[\mu_0 + 3\nu P]Q$	$\left[\mu_0 + \frac{32}{3\pi}\nu Q\right]Q$	$[\mu_0 + 2\nu Q]Q$	$[\mu_0 + 2\nu(2P + Q)]Q$
c_{12}	$-\frac{1}{\pi}\nu Q^2$	$-\frac{128}{45\pi^2}\nu Q^2$	$-\frac{4}{3\pi}\nu PQ$	$\frac{2}{3\pi}[\mu_0 + 2\nu(2P + 3Q)]_k P$
d_{12}	$\frac{3}{4}\nu(1 + k)Q^2$	$\frac{32}{15\pi}\nu Q^2$	νPQ	$-\frac{32}{5\pi^2}\nu_K PQ$
c_{12}	$-\frac{1}{\pi}\nu Q^2$	$-\frac{128}{45\pi^2}\nu Q^2$	$-\frac{4}{3\pi}\nu PQ$	$\frac{2}{3\pi}[\mu_0 + 2\nu(2P + 3Q)]_k P$
d_{12}	0	0	0	$\frac{32}{5\pi^2}\nu_K PQ$
c_{21}	$\frac{4}{3\pi}\nu PQ$	$\frac{128}{45\pi^2}\nu P^2$	$\frac{1}{\pi}\nu P^2$	$-\frac{4}{9\pi}\nu[5P - 6Q]_k P$
d_{21}	νPQ	$\frac{32}{15\pi}\nu P^2$	$\frac{3}{4}\nu(1 + 1/k)P^2$	$\frac{1}{3}\nu[P - 6Q]_k P$
c_{21}	$-\frac{4}{3\pi}\nu PQ$	$-\frac{128}{45\pi^2}\nu P^2$	$-\frac{1}{\pi}\nu P^2$	$-\frac{4}{9\pi}\nu[5P + 6Q]_k P$
d_{21}	0	0	0	$-\frac{1}{3}\nu[P - 6Q]_k P$
c_{30}	$-\frac{8}{15\pi}\nu P^2$	$-\frac{128}{225\pi^2}\nu P^2$	0	$-\frac{4}{3\pi}\left[\mu_{0K} - \frac{2}{5}\nu P\right]P$
d_{30}	0	0	0	0
c_{03}	0	$-\frac{128}{225\pi^2}\nu Q^2$	$-\frac{8}{15\pi}\nu Q^2$	$-\frac{8}{15\pi}\nu Q^2$
d_{03}	0	0	0	$\frac{8}{5}\nu_K PQ$

Intermodulation Products

Generated Modulation Voltages

From the foregoing results the voltage generated in a coil of N turns on a closed ferromagnetic core of cross-sectional area A can be found by the use of

$$e(t) = NA\ 10^{-8}\frac{dB}{dt}. \tag{52}$$

Components of this voltage segregated according to frequency are of the type

$$e_{mn}(t) = (mp + nq)NA\ 10^{-8}\left[c_{mn}\cos(mp + nq)t - d_{mn}\sin(mp + nq)t\right], \tag{53}$$

each proportional to its frequency and in general having two components in quadrature. The amplitudes of these will be designated by

$$E_{mn}' = (mp + nq)NA\ 10^{-8}\,c_{mn}, \tag{54}$$

$$E_{mn}'' = (mp + nq)NA\ 10^{-8}d_{mn}. \tag{55}$$

The amplitude of their resultant is

$$E_{mn} = (mp + nq)NA\ 10^{-8}\sqrt{c_{mn}^2 + d_{mn}^2}. \tag{56}$$

One component, if it greatly exceeds the other, may be taken as the generated distortion voltage. The various coefficients to which the components of the voltage are proportional have already been calculated, and also given in tabular or graphical form for specific instances. The relations

$$P = \frac{0.4\pi NI}{l}, \quad Q = \frac{0.4\pi NJ}{l},$$

where l is the mean length of magnetic path, may be used to convert these results into terms of the current amplitudes I and J. Where r-m-s quantities are used, they will be distinguished by bars over them.

Several features of the distortion are particularly outstanding. Perhaps chief among these is the dissimilarity of corresponding upper and lower side-frequency voltages. Inasmuch as these are products of a reactance modulator, they might be expected to be in the ratio of their frequencies, as they are found to be for one component. Often, however, a predominating component appears at each lower side frequency with no counterpart at the upper side frequency. This

component can be traced to the different axial slopes of the several branches of the loop, caused by their different points of origination. The slopes are fixed by the envelope of the magnetizing force; since this envelope is periodic in the difference of the fundamental frequencies, difference products will appear without corresponding sum products in the induction. Such a phenomenon is a fundamental property of the multivalued characteristic, and will occur wherever the envelope of a complex wave is instrumental in selecting the branch to be traversed.

No simple yet general rule seems to embrace the behavior of the various products in an iron core coil as governed by the amplitudes of the fundamental currents. Each voltage component is proportional to its frequency and to the product of two amplitudes, but these often enter in a complicated way. For fundamental frequencies close together all the higher order voltages vary directly with the hysteretic coefficient ν; for widely separated frequencies the distortion may depend also on the permeability through its effect on the axial slope of the minor loops. At the extremes of the amplitude ratio certain products or their components are found to be independent of one of the fundamental currents, the stronger one in some instances. When case 1 is applicable the third harmonic of the weaker fundamental current is suppressed below the value it would have without the stronger current superposed, while the third harmonic of the stronger fundamental is affected only slightly by the presence of a second frequency. Perusal of the table will disclose more detailed relations.

Distortion Measured in Coils

Voltages calculated by the theory have been compared with measured values for several coils using two common core materials. The agreement found provides a check of the theoretical predictions.

The two third order lower side frequencies of fundamental frequencies $p/2\pi = 760$ cycles per second and $q/2\pi = 600$ cycles per second are plotted in Fig. 4 for a higher frequency current of ten milliamperes in an iron dust coil of special design. The frequencies of the products are 920 and 440 cycles per second. These data were taken by I. E. Wood and the calculations were made by A. G. Tynan. These curves show the product as a function of the amplitude ratio more directly than the curves of Fig. 3.

Both upper and lower third order side frequencies have been measured by A. G. Landeen. The results are given in Figs. 5 and 6 for an annular core of iron dust. It is so wound that the magnetizing force is 0.04 times the current in milliamperes. The figures show two third

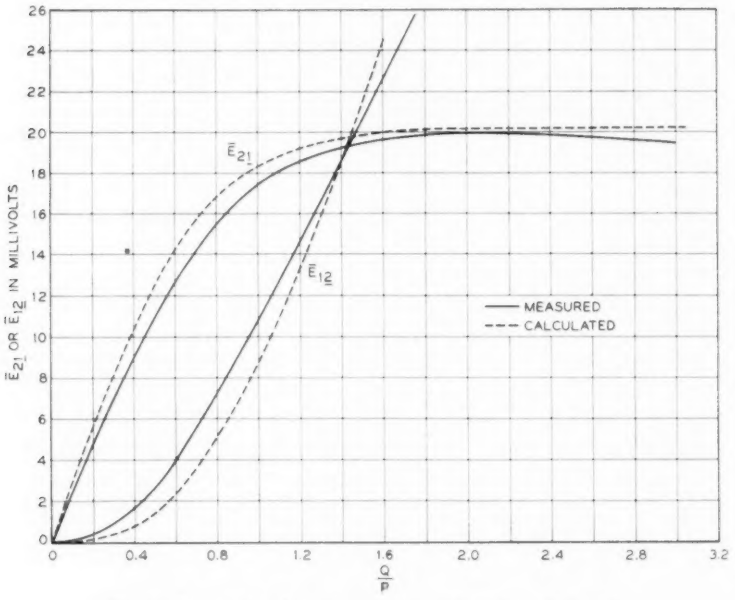


Fig. 4—Lower side-frequency voltages in an iron dust coil.

order products plotted against the fundamental current of higher frequency in each instance. For these measurements the current of one frequency was maintained at some fixed value and the amplitude of the other one varied. Both sum and difference products, but with different fundamental frequencies, are exhibited. The upper side

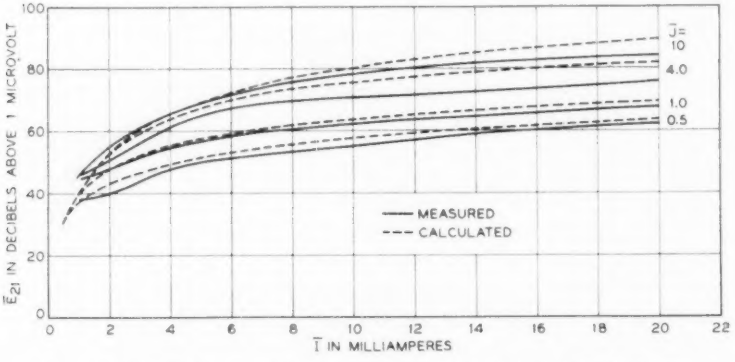


Fig. 5—Twenty-five kc. third order product of 9 and 7 kc., iron dust core. $\mu_0 = 24.5$, $p = 0.18$, $L_0 = 4.54$ mh., $H = 38.9\ I$.

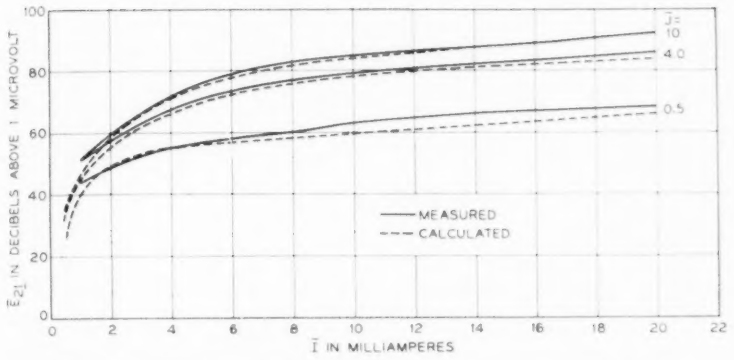


Fig. 6—Seventeen-kc. third order product of 13 and 9 kc., iron dust core. $\mu_0 = 24.5$, $\nu = 0.18$, $L_0 = 4.54$ mh., $H = 38.9\ I$.

frequency is the 25-kilocycle product of 9 and 7 kilocycles; the lower side frequency is the 17-kilocycle product of 13 and 9 kilocycles.

Some measurements to which subcase 1*b* is applicable are given in Figs. 7 and 8. Each curve gives a third order product for a permalloy dust core. The upper side frequency is 25 kilocycles generated by 9 and 7 kilocycle fundamental frequencies. The lower side frequency

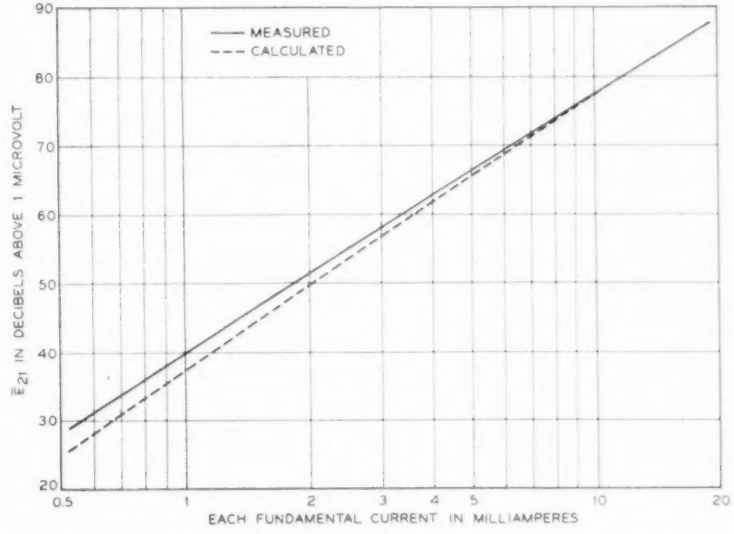


Fig. 7—Twenty-five kc. third order product of equal currents at 9 and 7 kc., permalloy dust core. $\mu_0 = 75$, $\nu = 0.41$, $L_0 = 5.45$ mh., $H = 36.0\ I$.

is 13 kilocycles from fundamental frequencies at 21 and 17 kilocycles. In these measurements both fundamental currents were changed simultaneously so as to be kept equal throughout. The approach to saturation at currents above ten milliamperes is apparent on these curves; below, the distortion voltage is proportional to the square of the current.

The measured curves are seen to agree well with the calculated ones in every instance, confirming theoretical values of the products within close limits. Eddy currents were negligible in all these coils because of the dust cores. The use of the formulae to determine important

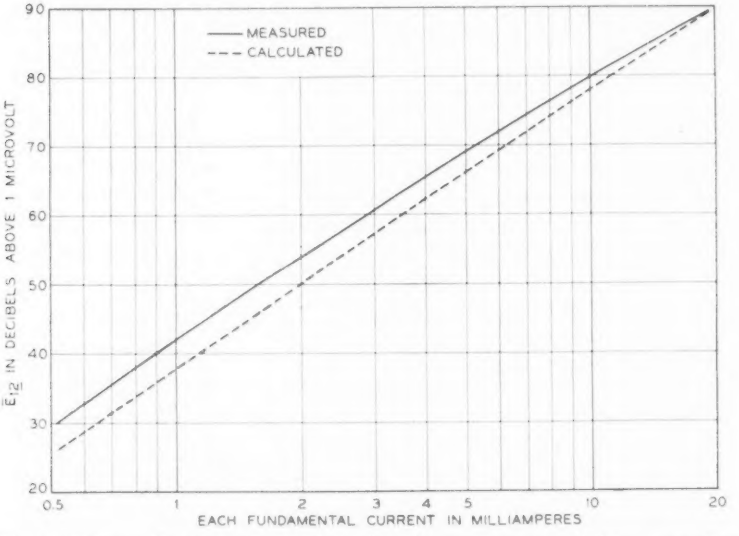


Fig. 8—Thirteen-kc. third order product of equal currents at 21 and 17 kc., permalloy dust core. $\mu_0 = 75$, $\nu = 0.41$, $L_0 = 5.45$ mh., $H = 36.0\ I$.

intermodulation products from the constants of the coils therefore seems to be justified.

Correlation with Single-Frequency Results

In the single-frequency case Peterson found the resistance and reactance at a fundamental frequency $\omega/2\pi$ to be increased by

$$\Delta R(\omega, H) = \frac{8}{3\pi}\omega L_0 \frac{\nu}{\mu_0} H,$$

$$\Delta X(\omega, H) = 2\omega L_0 \frac{\nu}{\mu_0} H,$$

respectively, on account of hysteresis and variable permeability, H being the maximum of the applied magnetizing force. The total non-linear reactance is then

$$X(\omega, H) = X_0 + \Delta X(\omega, H)$$

with

$$X_0 = \omega L_0,$$

the constant part, representing the reactance the coil would have if the permeability remained constant at its initial value.

The distortion voltages for a two-frequency input may be written in terms of these non-linear impedances. With some simplication this is done in Table II for special cases, using equation (56) and the relations

$$P = \frac{0.4\pi NI}{l}, \quad Q = \frac{0.4\pi NJ}{l}.$$

TABLE II

	Case 1a $\kappa \ll 1$ $k \ll 1$	Case 1b $\kappa \doteqdot 1$ $k \doteqdot 1$
E_{12}	$0.960\ \Delta R(p - 2q, Q)J$	$0.870\ \Delta R(p - 2q, Q)J$
E_{12}	$0.375\ \Delta R(p + 2q, Q)J$	$0.340\ \Delta R(p + 2q, Q)J$
E_{21}	$1.28\ \ \Delta R(2p - q, Q)I$	$0.870\ \Delta R(2p - q, P)I$
E_{21}	$0.500\ \Delta R(2p + q, Q)I$	$0.340\ \Delta R(2p + q, P)I$
E_{30}	$0.200\ \Delta R(3p, P)I$	$0.068\ \Delta R(3p, P)I$
E_{03}	—	$0.068\ \Delta R(3q, Q)J$
	Case 1c $k \gg 1$ $\kappa \gg 1$	Case 2 $\kappa \ll 1$ $k \gg \kappa$
E_{12}	$1.28\ \ \Delta R(p - 2q, P)J$	$0.212\ \kappa X(p - 2q, 2P + 3Q)I$
E_{12}	$0.500\ \Delta R(p + 2q, P)J$	$0.212\ \kappa X(p + 2q, 2P + 3Q)I$
E_{21}	$0.960\ \Delta R(2p - q, P)I$	$1.06\ \ \kappa\Delta R(2p - q, P - 0.944Q)I$
E_{21}	$0.375\ \Delta R(2p + q, P)I$	$1.23\ \ \kappa\Delta R(2p + q, P + 2.73Q)I$
E_{30}	—	$[0.425\ \kappa X_0(3p) - 0.200\ \Delta R(3p, P)]I$
E_{03}	$0.200\ \Delta R(3p, Q)J$	$0.200\ \Delta R(3p, Q)J$

The general formulæ of case 1 can be similarly represented, but not as concisely. Besides exhibiting the connection between intermodulation products and impedance changes, this table provides a convenient means for computing voltage components directly from data obtainable by single-frequency bridge measurements.

Hysteretic Impedances

The fundamental voltages are not included in the table. For them each component is separately significant. The singly primed E's are in phase with the corresponding magnetizing current and the doubly primed E's in quadrature. Hence, the former are resistance drops and the latter reactance drops, defining incremental components of impedance analogous to those mentioned for a single-frequency input.

For each of the fundamental frequencies the results developed previously may be used to determine these components. They represent the hysteretic resistance and the hysteretic reactance to the fundamental at hand, specified by a subscript p or q. They are tabulated in Table III for the special cases considered. The total resistance of the coil

TABLE III

	Case 1a $\kappa \ll 1$ $k \ll 1$	Case 1b $\kappa \doteq 1$ $k \doteq 1$	Case 1c $k \gg 1$ $\kappa \gg 1$	Case 2 $\kappa \ll 1$ $k \gg \kappa$
ΔR_p	$\frac{16}{15}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$	$\frac{256}{45\pi}\frac{N}{l}\frac{\nu}{\mu_0} pL_0J$	$\frac{8}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0J$	$\frac{8}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$
ΔX_p	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$	$\frac{64}{15}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$	$\frac{6\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0J$	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$
ΔR_q	$\frac{8}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0I$	$\frac{256}{45\pi}\frac{N}{l}\frac{\nu}{\mu_0} qL_0I$	$\frac{16}{15}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$	$\frac{16}{15}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$
ΔX_q	$\frac{6\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0I$	$\frac{64}{15}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0[2I + J]$

to either fundamental current can be calculated by adding to the value of ΔR from the table the resistance of the windings, the eddy current resistance, and the initial (viscosity) resistance, all evaluated for the frequency of the fundamental. The eddy currents must be so small that the flux density is substantially uniform across the cross section of the core. The reactance X_0 of the coil can be diminished by the eddy current reduction factor for the fundamental frequency and added to the hysteretic reactance to give the net reactance of the coil under these conditions.

The table is helpful in evaluating the effect of one fundamental current upon the other. Within the limits of the analysis, which in substance limits the permeability to linear variation with the field intensity, the hysteresis loss at any frequency is either increased or unchanged by the superposition of a second frequency. Nearly equal input currents whose frequencies do not differ greatly share equally the hysteresis loss. This amounts to about twice what it would if either fundamental flowed alone. If the frequencies differ greatly, the

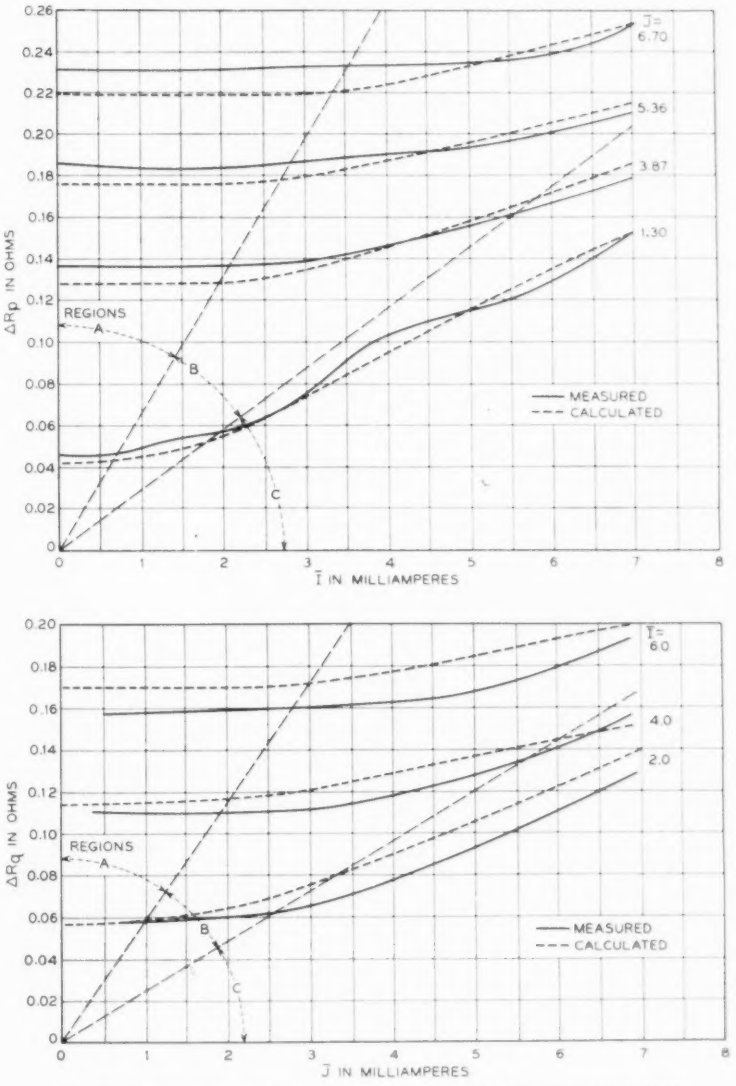


Fig. 9—Hysteretic resistance to one current in the presence of another.

loss to either is not affected by the other; if the amplitudes differ greatly, the loss to either is governed by the stronger current. In no case, at these weak fields, does the increased loss at one frequency

reduce the loss at another frequency, contrary to well established experiments at considerably higher fields, for which the hysteresis loss at one frequency may be reduced by superposing a magnetizing force at a different frequency. The inductance is the same to both funda-

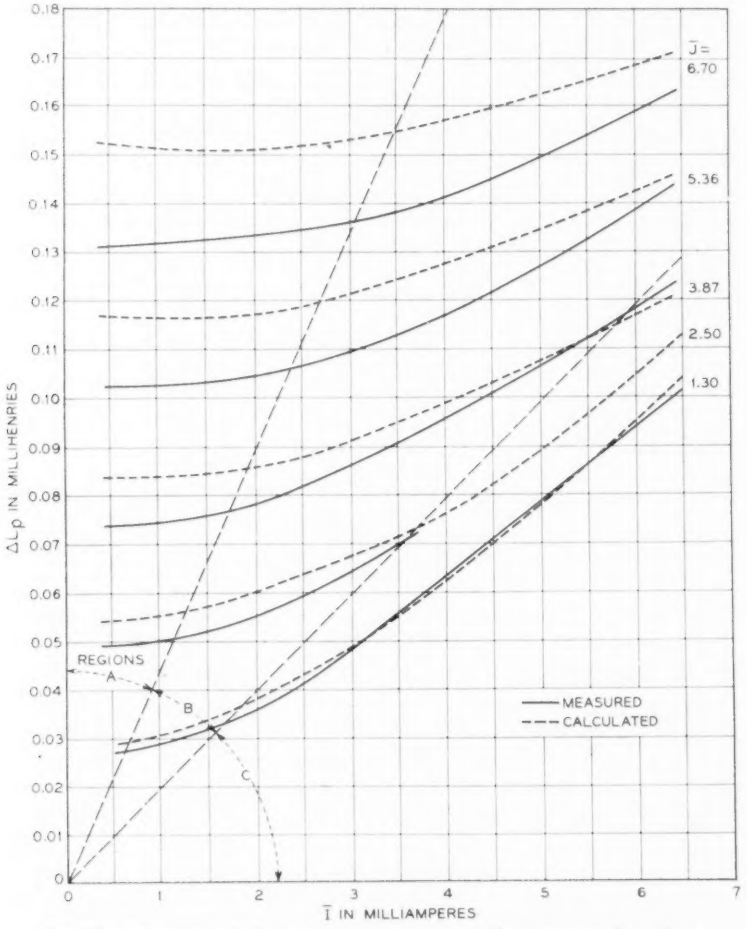


Fig. 10a—Hysteretic inductance to one current in the presence of another.

mentals in all cases except the second, for which a small difference exists. The effect of a superposed alternating current is always apparent through an increased inductance, although sometimes the increase may be slight; it is determined by the larger current.

The influence of one fundamental current upon another has been termed mutual crowding. Because of the increased attenuation, and at times because of resulting unbalance or phase shift, crowding becomes important when different frequencies or bands of frequencies

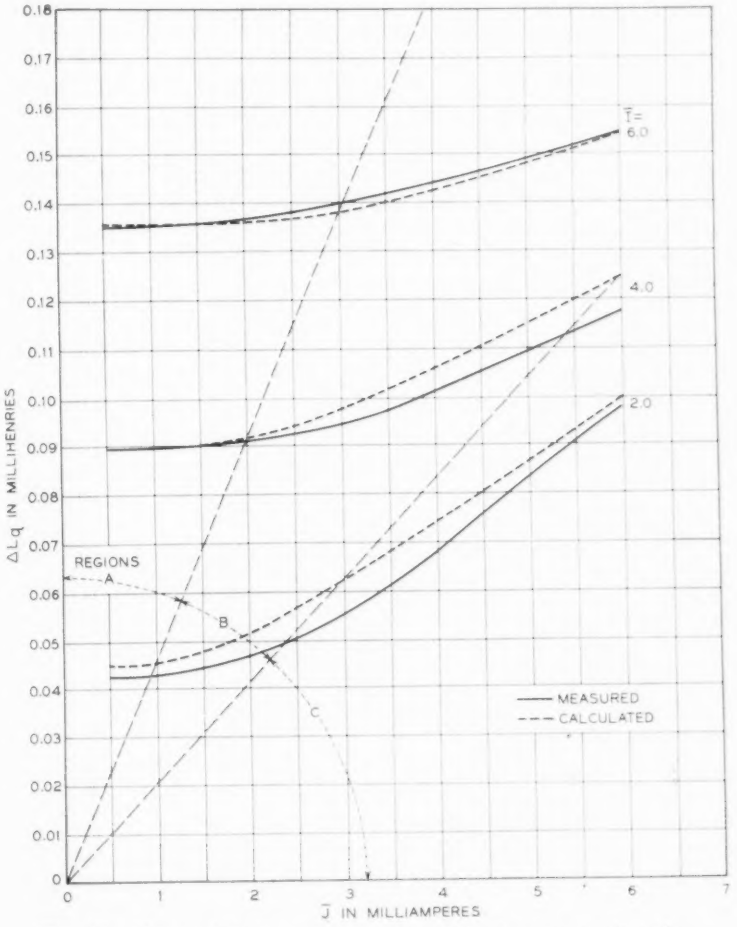


Fig. 10b—Hysteretic inductance to one current in the presence of another.

are transmitted simultaneously through a circuit including ferromagnetic material.

Incremental impedances for a twenty-two millihenry permalloy dust loading coil are given in Figs. 9 and 10. The current I had a frequency

of 550 cycles per second, and J 475 cycles per second. Measured and calculated values are plotted for comparison, with regions of applicability of the special subcases indicated. Portions of calculated curves falling in regions A or C were computed by the two sets of formulæ for subcases 1a and 1c, portions in the middle of region B by the formulæ for subcase 1b. The trend of the quantities measured is accurately portrayed by the calculations and agreement of the values is good.

All the curves commence at those values of resistance or inductance which would obtain in a single-frequency case with a current having the magnitude of the one here superposed in fixed amount. Upon increasing the variable current the measured quantities show an increase as it begins to preponderate, and eventually they approach asymptotically the values they would have if it flowed alone.

The measurements were made by L. R. Wrathall using a Maxwell inductance bridge with two imputs and a tuned detector. Eddy currents being of no consequence at the low frequencies employed, the chief sources of possible error are calibrations of the standards used and variation in the temperature of the coil during taking of the data. Changes in winding resistance caused by the latter are of the same order of magnitude as the changes in hysteretic resistance being observed. Precautions against both possibilities were taken.

Conclusion

The multiplicity of forms of complex hysteresis loops makes their analysis in general a complicated and difficult matter if indeed possible at all. Extensive experiments with two frequencies must be completed and the results classified according to the types of loops before an acceptable method of taking their form into account can be formulated. The parameters k and κ seem to be effective quantities for denoting concisely a particular form of loop in many instances.

A way of representing the behavior of complex loops more exactly than do Madelung's propositions is needed, and might be the fruit of precise experiments designed to clear up also the early closure and lack of closure which Lehde apparently found in minor loops. The tracing of complex loops is not simply cyclic, and only when a nearly complete magnetic cycle is executed between successive maxima in the magnetizing force can conditions approaching a cyclic state be expected to exist. Some experimental evidence of performance in other conditions is a present need which can perhaps be met by a thorough investigation of spiral characteristics. These seem to have been ignored entirely in the past, the literature dealing with sub-

sidiary loops only along a magnetization curve or a branch of a major loop.

Correlation of certain magnetic phenomena to a degree not heretofore attainable is made possible by the preceding development. Among at least some of these a qualitative connection has been well recognized. Flutter and allied effects are known to be aspects of modulation, and crowding is observed to accompany non-linear distortion quite generally. These features are related quantitatively by means of the theory, and are linked with their single-frequency counterparts. It thus becomes feasible to evaluate some of the more abstruse occurrences in terms of readily understandable effects simpler in nature; ultimate in this direction is the use of steady state results to forecast the behavior of transmitted speech or music.

Abstracts of Technical Articles from Bell System Sources

Reverberation Time and Absorption Measurements with the High Speed Level Recorder.[1] E. H. BEDELL and K. D. SWARTZEL, JR. It has been common in reverberation theories to neglect the effect of the stationary wave pattern in a room and to assume a logarithmic decay of the sound energy. In many cases this assumption of a constant decay rate is not fulfilled, and in particular it is well known that the decay curves as obtained with available instruments, which indicate either the pressure or the velocity at a point in the sound field, show marked fluctuations in the decay rate. The rate of decay, in general, varies during the decay period, from point to point in the room, and may depend upon the position of the sound source, and the location of absorbing materials. Very rapid fluctuations in the decay rate have commonly been averaged out by the measuring apparatus itself, either by making the indicating instrument sluggish in its action, or by measuring the average intensity over finite time intervals during the decay period. The slower, and perhaps more important, deviations from linearity in the decay curves have been either reduced, or averaged out, by a number of expedients. Among these are the use of rotating sound reflectors, or vanes, to break up the stationary wave pattern in the room; the use of frequency modulated, or warble, tones in place of a constant single frequency; moving the microphone to a number of positions in the room; and moving the sound source. Hunt has given some quantitative data on the effect of the warble tone, but similar data on other methods of smoothing out the decay curves are not available. This paper presents some data on the relative value of the above methods of improving the measured decay curves, and on the use of a motor driven rotating switch to connect in rapid sequence a number of microphones, placed in different parts of the room, into the measuring apparatus, for obtaining on a single curve a space average of the time decay pattern. Since many of the deviations from linearity in the decay curves have a "period" of 30 db or more, we should expect the accuracy of our values to be a function of the range through which the decay is measured, particularly when the range is not large compared to the period of the deviations. This effect is discussed for three values of the decay range, 30, 60 and 90 db.

[1] *Jour. Acous. Soc. Amer.*, January, 1935.

Standardization of Noise Meters.[2] R. G. McCURDY. A brief review of the present status of standardization of noise meters and measurements, and progress made to date by the technical committee on noise meters and noise levels of the American Standards Association.

A Rotating Mirror Oscilloscope.[3] R. F. MALLINA. When studying sound it is sometimes useful to project the wave-form of electrical or acoustical phenomena on a screen. A rotating mirror in combination with a vibrating mirror and a light source provide a convenient means of showing such waves. The problem of building an instrument for such a purpose is comparatively simple if a small screen is used in a dark chamber. However, when the screen is large enough to be viewed by a dozen or more persons, many difficulties arise.

The paper describes how the various parts of the apparatus may be coordinated in order to produce a comparatively bright, clearly defined wave with a small incandescent lamp in a room of average illumination. The vibrator used in the apparatus may be so constructed that its response is either inversely proportional to or independent of the frequency.

Shot Effect and Thermal Agitation in an Electron Current Limited by Space Charge.[4] G. L. PEARSON. The space current in a thermionic vacuum tube is not a steady flow of electricity but is subject to minute irregular fluctuations. The two most fundamental causes for these fluctuations are the random distribution of instants of emission of the individual electrons and the distribution of these electrons in velocity. The random emission produces shot noise which may be reduced by the space charge surrounding the cathode, while the velocity distribution produces thermal noise and is dependent upon the temperature of the cathode.

Although plausible theories of these effects have been given they have never been checked by accurate experiments because of the difficulties involved. By using two electrode tubes capable of producing a large space charge such measurements have now been made and are reported in this paper.

Simple Theory of the Three-Electrode Vacuum Tube.[5] H. A. PIDGEON. The physical principles upon which the operation of the three-element vacuum tube depends are presented in simple form and the terms

[2] *Elec. Engg.*, January, 1935; *Indus Standardization*, January, 1935.
[3] *Jour. S.M.P.E.*, December, 1934.
[4] *Physics*, January, 1935.
[5] *Jour. S.M.P.E.*, February, 1935.

usually applied to the tube, its operation as an amplifier, and a simple approximate method for computing the power output and percentage of distortion are explained.

No new material is presented in the paper although some of it is presented from a somewhat different point of view from that usually found in the literature. An effort has been made to present in reasonably compact form the essential features of the subject most useful to engineers interested in vacuum-tube applications.

The subjects discussed include: the portion of electron theory upon which the fundamental principles of vacuum-tube operation are based; space charge, the three-halves power law, temperature and voltage saturation; characteristics of the three-element tube; definition and physical significance of the terms plate resistance, transconductance, and amplification factor; dynamic characteristics, power output, and distortion; various means of coupling the vacuum tube to its associated circuits; and means for testing vacuum tubes for adequate thermionic emission.

Coaxial Communication Transmission Lines.[6] S. A. SCHELKUNOFF. A non-mathematical discussion of the mechanism whereby energy may be transmitted over long distances at high frequencies by the use of "coaxial conductors" is presented in this paper. A coaxial system consists of a cylindrical conducting tube within which a smaller conductor is coaxially placed. Such conductors, which reduce interference and crosstalk, are applicable for the transmission of telephone, telegraph, and television signals over a very wide range of frequencies.

Some Aspects of Quality Control.[7] W. A. SHEWHART. The object of this paper is to make clear what is meant by quality in a practical objective way that is subject to experimental verification and to consider some aspects of the problem of control. As a basis for judging the quality of current product it is necessary to obtain first of all adequate information, in the most efficient manner, on which to render a judgment. This can be accomplished by providing an inspection specification which is distinct from the design specification. One specifies the quantity and kind of evidence that is required as a basis for judging whether or not the quality of the product will attain its goal; the other specifies the goal. Certain elements of uncertainty must be allowed for in setting the goal. The discussion closes by pointing out the necessity of keeping a running report or record of the

[6] *Elec. Engg.*, December, 1934.
[7] *Mech. Engg.*, December, 1934.

evidence used in judging the quality of current product as a part of any scientific plan of making use of hindsight as well as foresight in controlling quality.

The Ionizing Effects of Meteors.[8] A. M. SKELLETT. It is shown that a meteor of average velocity has enough energy to cause ionization of atmospheric gases by impact. Recent experimental work by Frische and others on collisions of ions is interpreted as supporting the hypothesis that meteoric collisions do result in ionization. The afterglow of nitrogen is considered as a possible example of the process by which a meteor train remains glowing for a period of minutes and the coincidence of the region in which such trains are generally observed and of the E region of the upper atmosphere is pointed out. The spectra of bright meteors, while not showing atmospheric lines, are shown not to be inconsistent with the above hypothesis.

The behavior of the transatlantic short-wave radio telephone circuits of the American Telephone and Telegraph Company, during 1930, 1931, and 1932, is examined for possible meteoric effects. It is concluded that, in general, a rather large shower is necessary to affect them appreciably. This was to be expected since these circuits are normally under a continuous bombardment by random meteors. It seems possible that a certain degree of the variability (rapid fading, etc.) of received signals over such paths is due to this bombardment.

Results of radio pulse studies of the upper atmosphere, particularly by Schafer and Goodall, which are strongly suggestive of meteoric ionization, especially at times of special meteoric activity, are (1) sudden increases in ionization in the E region lasting for a period of minutes or less, and (2) increases of longer duration with maxima coincident in time with those of observed meteoric activity. Such tests made during the Leonid shower of November, 1932, were successful in correlating sudden increases in ionization in the E region with the visual observations of a number of bright meteors passing overhead. For the brightest meteor observed, the ionization increased to a value in excess of summer noon conditions.

It is pointed out that meteoric showers might take place in the F region which would be unobservable by ordinary visual means.

Taking into account the energy spent by the meteor in ionization, a mass for the brightest meteor, for which correlative data was obtained is roughly calculated to be 0.3 gram. Its estimated brightness was -1 magnitude.

The recombination coefficient at the height of the E region is calcu-

[8] *Proc. I.R.E.*, February, 1935.

lated from the rate of decrease of ionization after the passage of a meteor, to be less than 0.2×10^{-8} cubic centimeters per second.

Small Sapling Method of Evaluating Wood Preservatives.[9] R. E. WATERMAN and R. R. WILLIAMS. Permanence and toxicity are probably the most necessary characteristics of a wood preservative. Ease of injection, freedom from corrosive properties, cleanliness, cost, and the like are all important, but no material can be considered unless it displays a high degree of resistance to wood-destroying fungi and unless this toxic potency persists when the treated wood is exposed to the weather for long periods of time. The problem under discussion is that of appraisal of wood preservatives for these two characteristics within a reasonably short time.

In order to expedite tests of the permanency of pole preservatives, use is made of groups of small pine saplings treated with the preservative in question and set in the ground as miniature telephone poles. In these specimens weathering is relatively rapid on account of the large ratio of surface to volume, and poorly preserved material begins to decay in one or two years. Analyses and toxicity tests as well as observations of decay are made periodically. Seven years' experience indicates that the comparative preservative values of various salts, creosotes, oils, etc., may be estimated relatively cheaply, quickly, and with considerable reliability by this method.

A High Speed Level Recorder for Acoustic Measurements.[10] E. C. WENTE, E. H. BEDELL and K. D. SWARTZEL, JR. Two quite accurate means for recording rapid variations in sound intensity in a form suitable for visual inspection have been available for a number of years. One of these is the phonodeik, or one of its variants, and the other is a combination of a microphone and an oscillograph. When properly designed these devices record the actual wave form of the sound. However, for many acoustic measurements, a knowledge of the wave form is of secondary interest, whereas it is important that one should be able to record rapidly varying mean intensities over a wide range of values. From a record of the wave form it is not easy to determine the intensity with any degree of accuracy for a range greater than 20 or 30 db, but in some types of acoustic measurements it is highly desirable that the record cover a range of at least 60 db. Recently several types of instruments have been built which record, on a logarithmic scale, the mean power of the electrical input. These instruments, like that described here, may be used to plot the intensity

[9] *Indus. and Engg. Chem. (Analytical Edition)*, November 15, 1934.
[10] *Jour. Acous. Soc. Amer.*, January, 1935.

level in db as a continuous function of either time, frequency, or any other variable. The adaptability of such level recorders to acoustic measurements depends, among other factors, upon the range and accuracy of the logarithmic scale, and upon the effective speed of the recording mechanism. This recording speed is most conveniently expressed in terms of the rate, in db per second, at which the recorder is capable of following changes in the input power.

The level recorder described here consists essentially of an amplifier and rectifier, the output current of which is held at a substantially constant value automatically by a change in the gain of the amplifier, following changes in input power. The gain is varied by means of motor driven slide wire potentiometers graduated in logarithmic steps, the gain settings of which are recorded.

Some Applications of Modern Acoustic Apparatus.[11] S. K. WOLF and W. J. SETTE. Within the past two years there have been developed at the Bell Telephone Laboratories several electro-acoustic instruments designed to facilitate accurate measurement of a wide variety of acoustic phenomena. Three of these instruments are: an automatic level recorder, a crystal analyzer, and an acoustic spectrometer. Some of the types of acoustic studies for which these modern devices are well adapted may be of general interest and hence specific applications made at Electrical Research Products are described here. These include reverberation measurements, loud speaker response measurements, noise analyses, piano tone analyses, and studies on the singing voice. A brief description of the operating characteristics of the instruments is first given.

[11] *Jour. Acous. Soc. Amer.*, January, 1935.

Contributors to this Issue

ALFRED C. BECK, E.E., Rensselaer Polytechnic Institute, 1927. Test Department, New York Edison Company, summers 1926 & 27. Instructor in mathematics, Rensselaer Polytechnic Institute, 1927–28. Member, Technical Staff, Bell Telephone Laboratories, 1928–.

W. R. BENNETT, B.S., Oregon State College, 1925; A.M., Columbia University, 1928. Bell Telephone Laboratories, 1925–. Mr. Bennett has been engaged in the study of the electrical transmission problems of communication.

H. W. BODE, A.B., Ohio State University, 1924; M.A., 1926; Ph.D., Columbia University, 1935. Bell Telephone Laboratories, 1926–. Dr. Bode has been engaged in the study of transmission networks, such as wave filters, attenuation equalizers, and phase correctors.

R. P. BOOTH, S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1925. American Telephone and Telegraph Company, Department of Development and Research, 1925–34; Bell Telephone Laboratories, 1934–. Mr. Booth's work has had to do mainly with studies of crosstalk in open-wire and toll cable circuits.

EDMOND BRUCE, B.S., Massachusetts Institute of Technology, 1924. Radio service, U. S. Navy, 1917–19. Western Electric Company, 1924–25; Bell Telephone Laboratories, 1925–. Mr. Bruce has been engaged in the development of short-wave radio receivers and field-strength measuring equipment, and has specialized in directive antenna systems for short-wave radio communication. He was awarded the Morris Liebmann Memorial Prize by the Institute of Radio Engineers in 1932.

CHARLES R. BURROWS, B.S. in Electrical Engineering, University of Michigan, 1924; A.M., Columbia University, 1927. Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, Research Department, 1925–. Mr. Burrows has been associated continuously with radio research and is now in charge of a group investigating the propagation of ultra-short waves.

KARL K. DARROW, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of

Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

ALFRED DECINO, B.S., University of Colorado, 1928. Bell Telephone Laboratories, 1928–. Mr. Decino has been associated continuously with radio research and chiefly in studies of the propagation of radio waves.

R. L. DIETZOLD, S.B., Massachusetts Institute of Technology, 1925; Ph.B., Yale University, 1927; Department of Mathematics, Polytechnic Institute of Brooklyn, 1928–30. Bell Telephone Laboratories, 1930–. Mr. Dietzold has been concerned with research and design studies on transmission networks.

LOYD E. HUNT, A.B. in physics, Reed College, 1927. Wireless operator and ship wireless inspector, 1918–26. Radio Instructor, Oregon Institute of Technology, 1926–27. Director Radio-Electric Division, United Y.M.C.A. Schools, Seattle, Washington, 1927–29. Bell Telephone Laboratories, 1929–. Since coming with the Laboratories, Mr. Hunt has been employed in short-wave antenna development and is now engaged in ultra-short-wave propagation studies.

R. N. HUNTER, B.S., Worcester Polytechnic Institute, 1915. Test Department, General Electric Company, 1915–16. Research Assistant at Massachusetts Institute of Technology, 1916–18. American Telephone and Telegraph Company, Engineering Department, 1918–19; Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Hunter's work has been largely on problems of crosstalk reduction in open-wire and toll cable circuits.

ROBERT M. KALB, B.E.E., Ohio State University, 1927; Ohio State University, 1927–28. Bell Telephone Laboratories, 1928–. Mr. Kalb has been engaged in analyzing transmission problems, principally those of magnetic character.

H. P. LAWTHER, JR., B.A., University of Texas, 1912; M.A., Harvard, 1913; Ph.D., Harvard, 1916. Southwestern Bell Telephone Company, 1919–. Transmission and Protection Engineer, Texas Area, 1928–32; General Transmission and Protection Engineer, 1932–.